STANDARD REGRESSION VERSUS MULTILEVEL MODELING OF MULTISTAGE COMPLEX SURVEY DATA

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

In the Department of Community Health and Epidemiology

College of Medicine

University of Saskatchewan

By

MD ALOMGIR HOSSAIN

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the head of the Department or the dean of College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Request for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Department Head

Community Health and Epidemiology

College of Medicine

University of Saskatchewan

107 Wiggins Road

Saskatoon, Saskatchewan

S7N 5E5

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor Dr. Punam Pahwa for her motivation, intelligent supervising, consistent support, and encouragement. I greatly appreciate all her contributions of time, advice, thought, and patience during the entire process of completing my PhD thesis. I could not have imagined having a better supervisor for my PhD study.

I would like to thank to all of my committee members: Dr. June Hyun-Ja Lim, Dr. Bruce Reeder, Dr. Nazeem Muhajarine, Dr. Ivan Kelly, and Dr. Bonnie Janzen (chair) for their expertise, encouragement, and taking effort in reading my thesis and providing me the valuable suggestion in completing my PhD thesis.

My sincere thanks also go to Dean, the College of Medicine, University of Saskatchewan for offering scholarship during my PhD program.

I would like to thank to CIHR funded Canadian Heart Health Surveys New Emerging Team Grant for providing me the one year scholarship, computer and financial support to attend national and international conferences. I would specially like to thank the Canadian Centre for Health and Safety in Agriculture (CCHSA) for providing me the founding chair funding to attend several national and international conferences and own office space during my entire training period.

I would like to sincerely thank Dr. Bruce Reeder to allow me to use the Canadian Heart Health Surveys (CHHS) data, and Statistics Canada Research Data Centre (RDC), University of Saskatchewan, for providing me the access to use the National Population Health Survey (NPHS) data for my PhD thesis.

DEDICATION

This PhD thesis is dedicated to my family whose cooperation and sacrifices helped me to achieve my goal. Specially, my two daughters, Afra Nawar and Nazeefa Afreen missed me the most while I was busy with my PhD thesis work. I am also grateful to my father Golam Sarwar and mother Rashida Sarwar for their unconditional love and encouragement.

ABSTRACT

Complex surveys based on multistage design are commonly used to collect large population data. Stratification, clustering and unequal probability of the selection of individuals are the complexities of complex survey design. Statistical techniques such as the multilevel modeling – scaled weights technique and the standard regression – robust variance estimation technique are used to analyze the complex survey data. Both statistical techniques take into account the complexities of complex survey data but the ways are different.
This thesis compares the performance of the multilevel modeling – scaled weights and the standard regression – robust variance estimation technique based on analysis of the cross-sectional and the longitudinal complex survey data. Performance of these two techniques was examined by Monte Carlo simulation based on cross-sectional complex survey design.
A stratified, multistage probability sample design was used to select samples for the cross-sectional Canadian Heart Health Surveys (CHHS) conducted in ten Canadian provinces and for the longitudinal National Population Health Survey (NPHS).

Both statistical techniques (the multilevel modeling – scaled weights and the standard regression – robust variance estimation technique) were utilized to analyze CHHS and NPHS data sets. The outcome of interest was based on the question "Do you have any of the following long-term conditions that have been diagnosed by a health professional? – Diabetes".

For the cross-sectional CHHS, the results obtained from the proposed two statistical techniques were not consistent. However, the results based on analysis of the longitudinal NPHS data indicated that the performance of the standard regression – robust variance estimation technique might be better than the multilevel modeling – scaled weight technique

for analyzing longitudinal complex survey data. Finally, in order to arrive at a definitive conclusion, a Monte Carlo simulation was used to compare the performance of the multilevel modeling – scaled weights and the standard regression – robust variance estimation techniques . In the Monte Carlo simulation study, the data were generated randomly based on the Canadian Heart Health Survey data for Saskatchewan province. The total 100 and 1000 number of simulated data sets were generated and the sample size for each simulated data set was 1,731. The results of this Monte Carlo simulation study indicated that the performance of the multilevel modeling – scaled weights technique and the standard regression – robust variance estimation technique were comparable to analyze the cross-sectional complex survey data.

To conclude, both statistical techniques yield similar results when used to analyze the cross-sectional complex survey data, however standard regression-robust variance estimation technique might be preferred because it fully accounts for stratification, clustering and unequal probability of selection.

TABLE OF CONTENTS

PERMOSSION TO USE…….…..i

ACKNOWLEDGEMENTS……...……..…….ii

DEDICATION……..……..…….iii

ABSTRACT…….……..…..……iv

TABLE OF CONTENTS……..……..……..…..vi

CHAPTER 1 – INTRODUCTION…….…….……..1

1.1 Rationale……1

1.2 Study objectives…….……4

CHAPTER 2 – LITERATURE REVIEW……...5

2.1 Introduction……..……..5

2.2 Use of standard regression-robust variance estimation…….……10

2.2.1 Cross-sectional complex surveys……...….10

2.2.2 Longitudinal complex surveys……15

2.3 Use of multilevel modeling – scaled weights…….…...19

2.4 Monte Carlo simulation technique…….……...……25

2.5 Epidemiology of type 2 diabetes……..…….27

CHAPTER 3 – METHODS…….…….....30

3.1 Statistical methods to accomplish objective 1…….….....31

3.1.1 Standard regression for cross-sectional complex survey data…….…….31

3.1.1.1 Parameter estimation……..…..…..31

3.1.1.2 Variance-covariance estimation..…….…..33

3.1.1.2.1 Taylor linearization……34

3.1.1.2.2 Bootstrap variance estimation……35

3.1.2 Multilevel models for cross-sectional complex survey data……...36

3.1.2.1 Multilevel logistic random-intercept models……...37

3.1.2.2 Multilevel logistic random-coefficients models……..38

3.1.2.3 Parameter estimation for the multilevel model……...39

3.1.2.3.1 Maximum likelihood estimation (MLE)……39

3.1.2.3.2 Multilevel pseudo maximum likelihood (MPML)……....39

3.1.2.3.3 Adaptive quadrature……40

3.1.2.3.4 Relationship between regression coefficients obtained from multilevel modeling and standard regression……42

3.1.2.3.5 Scaling of weight……42

3.1.2.4 Variance estimation……...43

3.1.2.4.1 Sandwich estimator of the standard errors……...43

3.1.3 Goodness-of-fit test for logistic regression……...44

3.1.3.1 Goodness-of-fit test for survey sample……...45

3.2 Statistical methods to accomplish objective 2……48

3.2.1 Standard regression for longitudinal complex survey data……48

3.2.1.1 Marginal models for binary outcome……49

3.2.1.1.1 Generalized estimating equations (GEE)……51

3.2.1.1.2 Variance estimation……52

(i) Sandwich variance estimators……52

3.2.2 Multilevel modeling – scaled weights……53

3.2.2.1 Multilevel logistic random-intercept models……...…..54

3.2.2.2 Multilevel modeling random-coefficient models……...…...54

3.2.2.3 Multilevel pseudo maximum likelihood (MPML)……..…...55

3.2.2.4 Scaling of weight……55

3.2.2.5 Variance estimation……56

3.2.2.6 Goodness-of-fit……..56

3.3 Statistical methods to accomplish objective 3…….....58

3.3.1 Monte Carlo simulation Technique..……...……59

3.3.2 Monte Carlo simulation Technique using the CHHS……60

CHAPTER 4 –DESCRIPTION OF POPULATION OF STUDY……..…..……65

4 Introduction…….……65

4.1 Cross-sectional complex survey data: CHHS……..…….……65

4.1.1 Study design……..……..……66

4.1.2 Probability weight……..….……66

4.1.3 Study population ……...…….……68

4.1.4 CHHS data collection……..……..……..…….……68

4.1.5 Outcome variable of interest ……....…….…….69

4.1.6 Risk factors of type 2 diabetes……..…….…….……70

4.2 Longitudinal complex survey data: National Population Health

Survey (NPHS).,,,,,,,,,,,,,,,,,,,,,,,,……..……...72

4.2.1 Study design…….……..……..…….....72

4.2.2 Study population……..……..…….……75

4.2.3 NPHS data collection……..…….……..…….76

4.2.4 Probability weight……. 77

4.2.5 Outcome variable of interest……...77

4.2.6 Risk factors for type 2 diabetes……...……78

4.3 Simulated data for Monte Carlo simulation technique..……...80

4.3.1 Data set used for simulation technique……..…..82

4.3.2 Creation of 'weight' data file…….……..……82

4.3.3 Generating simulated data with the combinations of area, sex, age groups and PSU level…….83

4.3.4 Creating the simulated data sets with 100 and 1000 number of replications…..84

4.3.5 Creating final simulated data sets after linking each simulated data with weight file……..84

CHAPTER 5- ANALYSIS OF RESULTS ……88

5.1 Models for cross-sectional complex survey data…….…….88

5.1.1 Study population…….……...……..……...88

5.1.2 Descriptive analysis…….……...…..……90

5.1.3 Crude prevalence estimation……..…….……,……93

5.1.4 Modeling approach for cross-sectional complex survey data and results…...……..97

5.1.5 Comparison between the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique based on the CHHS……99

5.2 Models for longitudinal complex survey data…….…..……...103

5.2.1 Study population…….……...……...103

5.2.2 Descriptive analysis…….…….…….....104

5.2.3 Estimation of crude prevalence ……..…..……110

5.2.4 Modeling approach for longitudinal complex survey data…….……...114

5.2.5 Risk factors for type 2 diabetes based on the NPHS…….…….……...116

5.2.6 Comparison of the results obtained from the two technique……..…….……..117

5.3 Results based on Monte Carlo Simulation Technique…….…..…..121

5.4 Interpretation of results……...…….……...134

5.4.1 Interpretation of results based on the cross-sectional complex survey: CHHS……135

5.4.2 Interpretation of results based on the longitudinal complex survey: NPHS…….…….……..138

CHAPTER 6 – DISCUSSION AND CRITIQUE OF RESULTS ……..142

6.1 Introduction…….…….…….……..142

6.2 Objective 1: To compare the multilevel modeling-scaled weights technique and the standard regression-robust variance estimation technique by analyzing cross-sectional complex survey data...……..……..143

6.2.1 Prevalence of self-reported, physician-diagnosed type 2 diabetes and its risk factors among Canadians… ……...……..146

6.3 Objective 2: To compare the multilevel modeling-scaled weights technique and the standard regression-robust variance estimation technique by analyzing longitudinal complex survey data…….…….....……147

6.4 Objective 3: To investigate which statistical technique is optimal for analyzing complex survey data sets using a Monte Carlo simulation technique…...…….……..…..…….150

6.5 Strengths …….…….……..…….155

i) Data sets ……..…….…..……..155

ii) Analytical Technique …….……...……..156

6.6 Limitations……...…….…….…….…..……..156

i) Data sets ……..……..……156

ii) Analytical Technique ……..……..……..……...157

6.7 Future studies and recommendations……..…….....158

BIBLOGRAPHY……..…….....….....160

LIST OF TABLES

Table 3.3.1: The statistical formula for the assessment criteria…….……63

Table 4.1 Sample size stratified by province…….……69

Table 4.2 Longitudinal sample size stratified by province…….....75

Table 4.3 Response rate for each cycle…….……..….....76

Table 4.4 The estimated number of replications based on observed parameter estimates and their standard errors and expected accuracy…….……..….....81

Table 5.1 Number of participants with type 2 diabetic status in each Canadian province…….……..…….....89

Table 5.2 The number of participants in each covariate, stratified by self-reported type 2 diabetic status…….……..……..92

Table 5.3 Self-reported type 2 diabetes prevalence (95% C.I.) for all potential covariates included in the model.……..……..95

Table 5.4 Diabetes prevalence (%) stratified by type 2 diabetic status for each Province……..…..……...96

Table 5.5 Diabetes prevalence (%) stratified by type 2 diabetic status and location of residence for each Province…….....96

Table 5.6 Parameter estimates (standard errors) and their 95% confidence intervals based on the CHHS……..……101

Table 5.7 Distribution of self-reported, physician-diagnosed type 2 diabetic and non-diabetic participants (%), stratified by cycles…….....104

Table 5.8 Number (%) of self-reported, physician-diagnosed type 2 diabetic cases according to the potential risk factors, based on Cycle 1 (1994 – 95)…..……..107

Table 5.9 Distribution of participants (%) stratified by cycles and provinces…….……108

Table 5.10 Prevalence of type 2 diabetes (95% confidence interval) stratified by cycles……..…….…..……..109

Table 5.11 Self-reported type 2 diabetes prevalence (95% confidence interval) for potential covariates included in the final model……...……112

Table 5.12 Estimates (Standard Errors) and their 95% confidence intervals based on the NPHS…….……119

Table 5.13 Results for assessment criteria to compare the performance of the multilevel modeling-scaled weights technique and the standard regression-robust variance estimation technique based on Monte Carlo simulation ……..…….…….127

Table 5.14.1 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 1)…… ……..…….132

Table 5.14.1 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 2)…… ..……132

Table 5.14.1 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 3)…… …….……..133

Table 5.14.1 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 4).……..……..133

Table 5.15 Odd ratios (OR) and 95% confidence intervals (95% C.I.) based on the standard regression-robust variance estimation technique using the CHHS …...……..137

Table 5.16 Calculation of odd ratios for interaction term based on CHHS …….…….....138

Table 5.17 Odd ratios (95% confidence intervals) based on the standard regression–robust variance estimation technique using the NPHS…….…….....140

Table 5.18 Calculation of odd ratios for interaction term based on NPHS……..……..……..……141

LIST OF FIGURES

Figure 3.1 Flow chart of statistical methods used to accomplish Objective 1…..…… …..47

Figure 3.2 Flow chart of statistical methods to accomplish Objective 2 …….…….....……..57

Figure 4.1 The survey methodology for NPHS data……...……...74

Figure 4.2 Data linkage between simulated data and weight file created from Saskatchewan data……..…….....85

Figure 5.1 Distribution of self-reported, physician-diagnosed type 2 diabetes among Canadian provinces ..……..……90

Figure 5.2 Prevalence of self-reported, physician-diagnosed type 2 diabetes over time…….……..……...109

Figure 5.3 Comparison between multilevel modeling-scaled weights technique and standard regression-robust variance estimation technique based on the results obtained from the analysis of simulated data with two sample sizes ……..…….130

LIST OF ABBREVIATIONS

CHHS: Canadian Heart Health Survey

NPHS: National Population Health Survey

GEE: Generalized Estimating Equations

SR-RV: Standard regression-robust variance

MM-SW: Multilevel modeling-scaled weights

CHAPTER 1

INTRODUCTION

1.1 Rationale

Complex surveys based on multistage design are frequently used to conduct large population studies. Stratification, clustering and unequal probability of the selection of individuals are the complexities of complex survey design that are also known as design effects. Sampling units may not be independent because of stratification and clustering in complex surveys. Special statistical techniques are required to analyze data obtained from complex surveys to take into account the complexities associated with such survey design. Statistical analysis conducted without taking into account the characteristics of longitudinal data, such as within-subject correlation due to repeated measurements and design effects of complex survey design, may lead to bias and invalid parameter estimates and standard errors [1, 2]. The selection of sample units from a finite population and the processing of responses and measurements are part of complex survey design. The modeling of the variation of data due to these processes in complex surveys is part of the inferential process [3]. Several studies have indicated that parameter estimates could be inconsistent without taking into account the design effects of complex surveys [2, 4-7].

Population-based cross-sectional and longitudinal complex surveys are commonly conducted to collect huge amounts of information on various health outcomes, such as chronic conditions and the associated risk factors. Standard statistical models have been developed based on the assumption of simple random sampling. In complex survey design, since sampling units are not independent, standard errors, confidence intervals and p-values

obtained from the standard approach will be invalid because of the lack of independent observations [2, 8, 9].

A number of statistical methods are proposed to analyze cross-sectional and longitudinal complex survey data for continuous and discrete outcomes. The multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique (e.g., Taylor linearization, jackknifing, and bootstrapping) are the most frequently used techniques for analyzing data obtained from cross-sectional and longitudinal complex surveys. Both the multilevel modeling–scaled weights (MM-SW) technique and the standard regression–robust variance (SR-RV) estimation technique take into account the complexities of complex survey design, but the ways of taking these design effects into account are different. In contrast to cross-sectional complex surveys, longitudinal complex surveys have an additional characteristic—within-subject correlation due to repeated measurements on each individual. This additional feature makes the statistical analysis of longitudinal complex survey data more difficult compared with cross-sectional complex survey data. The statistical analysis of longitudinal complex survey data must take into account the within-subject correlation characteristics of repeated measurements in addition to stratification, clustering and unequal probability of selection.

A Medline search revealed a few studies that attempted to compare the MM-SW technique and the SR-RV estimation technique based on multistage complex survey datasets [10-13]. However, a definite conclusion about which technique is preferable for analyzing complex survey data was not reached in those studies.

The overall goal of this thesis is to conduct a comparison between the multilevel modeling–scaled weights technique and the standard regression–robust variance (such as

bootstrapping) estimation technique. The similarities and differences between these statistical methods will be explored by applying the proposed techniques to analyze real-life data obtained from cross-sectional and longitudinal complex surveys.

It is important to establish the properties of statistical methods so that researchers and statistical analysts can use it with confidence. Real-life survey data rarely satisfy all the assumptions required to use most statistical methods. Simulation is a great technique to determine the power of statistical methods. Simulation techniques are used in almost half of the articles published in the Journal of the American Statistical Association [14]. Today, simulation is a less problematic way to test the power of statistical methods because of the availability of computer software.

In this thesis, the Canadian Heart Health Survey (CHHS) (a complex cross-sectional survey) and the National Population Health Survey (NPHS) (a complex longitudinal survey) datasets will be used to accomplish our objectives. The CHHS and NPHS datasets are unique datasets because results based on these datasets can be generalized to the entire Canadian population. Monte Carlo simulations are conducted to compare the MM-SW approach and the SR-RV estimation approach based on cross-sectional complex survey data sets.

The outcome of interest for application of the proposed statistical methods is type 2 diabetes. Type 2 diabetes is a complex chronic disease, and the etiology of type 2 diabetes is not yet completely understood. The prevalence of type 2 diabetes is increasing rapidly all over the world [15] . Indeed, the expected prevalence of type 2 diabetes is 2.4 million by the year 2016 in Canada [16]. One of the main causes of cardiovascular disease (CVD), blindness, heart disease and kidney failure is type 2 diabetes [17].

1.2 Study objectives

Objective 1:

To compare the use of the multilevel modeling–scaled weights (MM-SW) technique and the standard regression–robust variance(SR-RV) estimation technique to analyze cross-sectional complex survey data.

Objective 2:

To compare the use of the multilevel modeling–scaled weights (MM-SW) technique and the standard regression–robust variance (SR-RV) estimation technique in analyzing longitudinal complex survey data.

Objective 3:

To investigate which statistical method is optimal for analyzing cross-sectional complex survey data using Monte Carlo simulation techniques.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The statistical analysis of survey data depends on the characteristics of the sampling design. Simple random sampling (SRS) is a standard sampling design in which individuals are assumed to be independent, and each individual has an equal probability of selection [18]. SRS is not a preferred sampling design for conducting large population surveys for several reasons [19]. It is financially expensive to conduct large surveys based on SRS, and such surveys require a longer time to collect data than do multistage complex surveys. A multistage complex survey design involves stratification, clustering and unequal probability of selection of sampling units.

Population-based large health surveys frequently use multistage complex survey design. There are several reasons to conduct multistage complex surveys: they are economical, and they make it easy for interviewers to collect information. There are some disadvantages to complex survey design, which are mainly related to the statistical analysis of the data obtained from the survey. The sampling units might be correlated within a cluster, and the probability of selection of all of the sampling units might not be equal. These features of complex survey design, such as stratification, clustering and unequal probability of selection make significant impact in the estimation process. The parameter estimates will be invalid if these features of complex survey design are ignored, and the statistical inference based on such invalid parameter estimates will be erroneous [101].

In the last few decades, several statistical methods have been developed to analyze complex survey data. However, the most commonly used methods to analyze complex survey

data are the multilevel modeling–scaled weights (MM-SW) technique and the standard regression–robust variance (SR-RV) estimation technique.

Few studies have made the comparison between using standard regression and multilevel modeling techniques to analyze complex survey data [10, 12, 20-24]. Multilevel models are also referred to as mixed-effects or random-effects models, random-coefficient models or hierarchical models. A study was conducted by Moerbeek et al to compare between traditional methods used for regression analysis and multilevel models based on the analysis of multicenter intervention studies for continuous outcomes [12]. The comparison was made based on the estimated regression coefficients and their standard errors. The authors found that the standard errors of the regression coefficients were underestimated using traditional regression methods. Because of the smaller standard errors, the confidence intervals became narrow, and there was a higher possibility of making a type I error. The authors preferred multilevel models over traditional methods (i.e. ordinary logistic regression for binary outcome). They also observed that the magnitude of the regression coefficients and their standard errors were affected by using a multilevel modeling approach for the data of an unbalanced design. Several statistical methods have been proposed to analyze longitudinal binary data. Most of the methods can be divided into two groups: (i) subject-specific (SS) models and (ii) population-averaged (PA) models. The generalized estimating equation (GEE) approach (Liang and Zeger) is most commonly used to fit PA models and multilevel models based on pseudo maximum likelihood algorithm used to fit random-effects or SS models[14].

Standard logistic regression provides biased standard errors (SE) for analyzing the longitudinal data because it violates the independence assumption. If regression models ignore the dependency of the observations within subjects, then such models tend to overestimate the

standard errors of time-varying covariates and underestimate the standard errors of time-invariant covariates [11]. Neuhaus et al. compared a cluster-specific model (i.e., mixed-effect logistic model) and a population-averaged model (i.e., GEE) for analyzing correlated binary data[10]. Clustering may be due to repeated measurements within-subjects or may be due to sub-sampling of a primary sampling unit (PSU). The authors showed that the regression coefficients obtained from a mixed-effect logistic model were higher than those from a population-averaged model. It was also shown by Liang and Zeger that the regression coefficients obtained by using a random-effects model were higher than the regression coefficients obtained by using a population-averaged model [38]. Liang and Zeger also showed that there was a mathematical relationship between these two types of regression coefficients [38] .

Marginal models and random-effects models were frequently used to analyze longitudinal complex survey data with binary outcomes in epidemiology. Corriere and Bouyer discussed how to choose statistical methods based on the analysis of longitudinal binary data [21]. The results from the analysis of longitudinal binary data indicated that there were substantial differences in the parameter estimates from random-effects models and marginal models. The inter-individual heterogeneity was the main reason for the differences between the estimates of these two methods. The authors also pointed out that the choice of a model to analyze the longitudinal data depends on the research objective. If the research objective is to determine the association between the populations mean of the outcome over time and the risk factors, then a marginal model is appropriate. If the objective is to study individual risk factors for etiological consideration, then the random-effects model is appropriate because this method adjusts for the non-observable individual characteristics [25]. After comparing the two

methods, the authors recommended that random-effects models were more suitable than marginal models for analyzing longitudinal binary data.

A large number of simulation studies were conducted by Rodriguez and Goldman to assess the estimation procedures for multilevel models with binary outcomes[23]. The results of these simulation studies specified that the estimated fixed effects and the variance components would be biased if the random effects were sufficiently large and if the number of observations within a given level of clustering was small . Rodriguez and Goldman also found that the fixed effect estimates were similar between standard logit models and multilevel logit models if the hierarchical structure of the data was ignored [23]. Finally, the authors anticipated that an alternative estimation procedure would be required for handling hierarchical data with binary outcomes. In a random-intercept logistic model, the interdependencies among the repeated observations within-subjects were explicitly taken into account [23]. The absolute values of the estimates obtained from random-effects models were generally larger than those obtained from GEE models. These differences between the GEE and random-effects models depend on the correlation between the repeated measures. Frank B. Hu et al also suggested that the selection of statistical methods to analyze longitudinal complex survey data should depend on the research objective [22]. The GEE approach is preferable compare to the random-effects models if the research objective involves group differences, while the random-effects models are preferable when the research objectives involve determining the change in individual responses. The GEE approach provides robust variance estimation, whereas the random-effects approach may be sensitive to different assumptions about the variance and covariance structure [22]. A comparison was explored between marginal and mixed-effects models based on an analysis of human papillomavirus

(HPV) natural history data. Xue et al. found that the parameter estimates obtained using a mixed-effects model was higher than those obtained using marginal models [20]. The standard errors of the estimated regression coefficients were also higher in the mixed-effects model, but the significance levels obtained were similar in both types of models. Some disadvantages, such as being computationally intensive and more likely to have problems with convergence, are found in the mixed-effect model compared with the marginal model. Therefore, marginal models are sometime preferred for analyzing data obtained from epidemiological studies. Kuchibhatla and Fillenbaum compared random-intercept models and marginal (GEE) models based on an analysis of longitudinal data with binary outcomes [24]. Both statistical methods were used to analyze longitudinal binary data- their findings indicated that the estimated regression coefficients and their standard errors obtained from random-intercept models were larger than those obtained from marginal (GEE) models. The differences in the estimates from random-intercept models and GEE models are due to correlations between the repeated observations. The authors did not make any comment regarding which method was better for analyzing longitudinal data, but they concluded that the marginal (GEE) model was appropriate when the research objective was to investigate the between-subject effects and the random-intercept model was appropriate when the research objective was to investigate subject-specific effects.

A simulation study is the best way to assess the performance of two statistical methods. Masaoud and Stryhn conducted a simulation study to compare the performances of random-effects models and marginal (GEE) models to analyze binary repeated measurements data [26]. The results based on the analysis of the simulated data using random-effects models showed that the parameter estimates were biased when autocorrelation was present in the data,

while the marginal models provided estimates close to the marginal parameters. A number of studies have been conducted to compare multilevel models/subject-specific models/random-effects and population-averaged models, but most of the studies did not consider the effects of the sampling design characteristics at their analysis stage.

A literature review related to the multilevel modeling–scaled weights (MM-SW) technique and the standard regression–robust variance (SR-RV) estimation technique is discussed in sections 2.2 and 2.3, respectively. The literature review related to simulation studies is discussed in section 2.4. Both the MM-SW technique and the SR-RV estimation technique were applied to real-life cross-sectional and longitudinal complex survey data to compare these two methods. The significant risk factors for type 2 diabetes among rural and urban populations in Canada were determined based on these two methods. A literature review related to the epidemiology of type 2 diabetes is given in section 2.5.

2.2 Use of standard regression–robust variance estimation

2.2.1 Cross-sectional complex surveys

In cross-sectional complex surveys, sampling units are measured at single time points. Standard regression techniques are generally used to analyze cross-sectional survey data in which the response variable can be continuous, categorical or count [27]. Logistic regression models introduced by McFadden are widely used to analyze the data for binary responses [107]. Let $Y = (y_1, y_2, ..., y_n)$ be a vector of response variable and $\vec{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ be a vector of explanatory variable where i=1,2,3,….,n. Assume the response variable y_i is dichotomous with a value of 1 or 0, where "1" means success and "0" means failure. The probability density function of y_i is

$$f(y_i \mid \theta) = \mu_i^{\,y_i}[1-\mu_i]^{1-y_i} \tag{2.1}$$

where θ is the parameter vector and $\mu_i = pr(y_i = 1 \mid \vec{x}_i)$ [27].

In general, the mathematical form of the logistic regression is

$$\text{logit}\left(pr(y_i = 1 \mid \vec{x}_i)\right) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i\,j} \tag{2.2}$$

where the regression coefficients $\vec{\beta} = \left(\beta_0, \beta_1, \beta_2, ..., \beta_p\right)$ need to be estimated. Based on Equation (2.1), the log-likelihood function is written as

$$L\left(\vec{\beta}\right) = \sum_{i \in S} y_i \ln F(\vec{x}_i \vec{\beta}) + \sum_{i \in S} (1-y_i) \ln\left\{1 - F(\vec{x}_i \vec{\beta})\right\} \tag{2.3}$$

where $\mu_i = F\left(\vec{x}_i \vec{\beta}\right) = \dfrac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$ and x_{ij} is the j^{th} covariate (j=1,2,3,..., p) for the i^{th} subject. The regression coefficient can be estimated by the maximum likelihood estimation technique, i.e., the solutions of the score equations $\dfrac{\partial L\left(\vec{\beta}\right)}{\partial \vec{\beta}} = 0$ will provide the regression coefficients $\vec{\beta}$.

The variance-covariance matrix of the estimated regression coefficients is obtained using the Fisher information matrix [36]:

$$\mathrm{I}\left(\vec{\beta}\right) = -E\left(\frac{\partial^2 L}{\partial \vec{\beta}^2}\right), \text{ i.e., } Cov\left(\vec{\beta}\right) = \mathrm{I}^{-1}\left(\vec{\beta}\right).$$

There are two basic disadvantages of using the above procedure to estimate the parameters and the variance-covariance matrix for complex survey data:

(i) It does not take into account the unequal probability of selection.

(ii) It does not take into account the stratification and clustering of complex survey data for variance estimation.

If the probability of selection for each individual is not equal, then the above procedure is not appropriate for producing valid parameter estimates. In order to obtain the unbiased parameter estimates, we have to take into account the unequal probability of selection [25, 28]. The appropriate sample weight should be used to take into account the unequal probability of selection as well as any non-responses to analyze complex survey data. The corresponding log pseudo-likelihood function using sampling weight ω_i [101] is

$$L_w(\vec{\beta}) = \sum_{i \in S} \omega_i y_i \ln F(\vec{x}_i \vec{\beta}) + \sum_{i \in S} \omega_i (1 - y_i) \ln\left(1 - F(\vec{x}_i \vec{\beta})\right) \tag{2.4}$$

where $F(\vec{x}_i \vec{\beta}) = \dfrac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}$ and S is the set of all observed observations.

The regression coefficients and their variance-covariance estimators can be obtained from the score equations $\dfrac{\partial L_w(\vec{\beta})}{\partial \vec{\beta}} = 0$ and $Cov(\vec{\beta})_{\omega} = \mathrm{I}_w^{-1}(\vec{\beta})$, where $\mathrm{I}_w(\vec{\beta}) = -E\left(\dfrac{\partial^2 L_w(\vec{\beta})}{\partial \vec{\beta}}\right)$.

It is not sufficient to account for the unequal probability of selection to estimate the valid variance components in analyzing complex survey data. Clustering and stratification in complex surveys should also be taken into account because the sampling units may be correlated to each other [30]. A consistent variance-covariance matrix of parameter estimates can be obtained using Taylor expansion of $G(\vec{\beta})$ at $\vec{\beta} = \hat{\vec{\beta}}$, where $G(\vec{\beta}) = \dfrac{\partial L_w}{\partial \vec{\beta}}$ [30]. Binder

proposed the theory of variance estimation using Taylor expansion for model parameter estimates based on complex survey data [30]. The variance estimator of $\hat{\vec{\beta}}$ is

$$V\left(\hat{\vec{\beta}}\right)=\left[\left\{\frac{\partial G\left(\vec{\beta}\right)}{\partial\vec{\beta}}\right\}^{-1}V\left(\vec{\beta}\right)\left\{\frac{\partial G\left(\vec{\beta}\right)}{\partial\vec{\beta}}\right\}^{-T}\right] \tag{2.5}$$

The square roots of the diagonal elements of $V\left(\hat{\vec{\beta}}\right)$ are called the "robust standard errors" [31].

Variance estimators of parameter estimates are important to determine the quality of estimation of population parameters, and standard errors of parameter estimates can be obtained from the variance estimators. Standard errors of parameter estimates are used to determine the confidence intervals of parameter estimates. In complex survey data, it is complicated to estimate the valid sampling variance because of complexities such as clustering, stratification and unequal probability of selection of complex survey data. Analytical methods such as the Taylor linearization method and re-sampling methods such as jackknifing, balanced repeated replication (BRR), and bootstrapping are the main techniques used to estimate the variance estimators of parameter estimates. Re-sampling methods are easier to apply to complex survey data than the Taylor linearization method for determining the standard errors [28]. In the Taylor linearization method, the computation of the partial derivative of the log-likelihood function for certain parameters might be difficult [28, 32]. Re-sampling methods are often used to estimate the variance estimators of parameters. The jackknife re-sampling approach has fewer computational problems compared with the linearization method. The BRR re-sampling method provides consistently better variance estimation of parameters than do jackknifing and the linearization method [32].

Among all the re-sampling methods (jackknifing, bootstrapping and BRR), bootstrap re-sampling offered the best estimation of standard errors after taking into account the design features of complex survey data [7]. The bootstrap re-sampling technique was first developed by Efron for independent and identically distributed (iid) data [33]. Rao and Wu proposed an extension of the bootstrap technique for complex survey data [28]. Rao, Wu and Yue modified the bootstrap technique for complex survey data so that it can take into account the design features and unequal probability of selection [34]. This modification of the bootstrap technique involved the scale adjustment of the survey weights [35]. The current bootstrap re-sampling technique takes into account the effect of design features (stratification and clustering) and weight adjustments [7]. A bootstrap variance estimation technique was also proposed for multilevel modeling using the Rao and Wu bootstrap technique [36].

The bootstrap re-sampling method generates artificial data sets of the same size and structure as the original data set. Let $\hat{\bar{\beta}}_b^*$ be the parameter estimator for both artificial data sets where b=1,2,3 ….B.

The bootstrap variance estimator of $\hat{\bar{\beta}}$ is defined by

$$\hat{V}_{BS}\left(\hat{\bar{\beta}}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\bar{\beta}}_b^* - \hat{\bar{\beta}}^*\right)^2 \qquad (2.6)$$

where $\hat{\bar{\beta}}^* = \frac{1}{B}\sum_{b=1}^{B}\hat{\bar{\beta}}_b^*$ and B is the number of repetitions.

2.2.2 Longitudinal Complex Surveys

In longitudinal complex survey data, sampling units are measured repeatedly over time. Analyses of longitudinal data are complicated compared to cross-sectional data because the repeated measurements obtained for a given subject are correlated. The estimation of parameters will be biased if the within-subject correlations are ignored.

Generalized linear models (GLMs) were first introduced by Nelder and Wedderburn to fit observed data in which the distribution of outcome variables belongs to an exponential family (e.g., normal, binomial, poisson and gamma) [18, 37]. The regression coefficients in GLMs are obtained from the maximum likelihood (ML) method [18, 38]. The distribution of outcome variables is necessary to determine the maximum likelihood (ML) function. The quasi-likelihood function, introduced by Wedderburn , is an alternative to the ML function in which it is not necessary to specify the distribution of outcome variables [37]. It requires only the relationship between the mean and the variance of the outcome [1, 37, 38].

Liang and Zeger proposed the generalized estimating equations (GEE) as an extension of generalized linear models to analyze longitudinal data [14, 38, 39].

The GEE approach is used mainly for marginal models based on the quasi-likelihood theory, which was introduced by Wedderburn [40]. Quasi-likelihood and pseudo likelihood are not the same function [40]. Bahadur first proposed the marginal model for discrete data that takes into account within-subject correlation based on likelihood inference [1, 38]. The GEE based on marginal models is widely used to analyze longitudinal data because it is not necessary to know the distribution of the outcome variable and it takes into account the within-subject correlations. The GEE is also used to analyze the clustered data, which takes into account the intra-cluster correlation, but it may provide overstated type I errors [41]. The

advantage of GEE is that the regression coefficient estimates are consistent and efficient, even though the within-subject correlation structure is specified incorrectly [14, 38, 39].

Suppose $Y_i = (y_{i1}, y_{i2}, ..., y_{ir})^T$ is a $r \times 1$ vector of dichotomous response for the i[th] subject (i=1, 2,…, n) where r indicate the number of repeated measurements within i[th] subject and $\mu_i = (\mu_{i1}, ..., \mu_{ir})^T$ denotes the mean vector for the i[th] subject. The GEE is an alternative approach to standard likelihood equations for estimating parameter estimators [1, 38]. The estimator $\hat{\vec{\beta}}$ of $\vec{\beta}$ can be obtained by solving the following set of score equations [1, 19, 39]:

$$U\left(\hat{\vec{\beta}}\right) = \sum_{i=1}^{n} D_i^T V_i^{-1} (Y_i - \mu_i) = 0$$

$$= \sum_{i=1}^{n} D_i^T \left(A_i^{1/2} \Re_i(\alpha) A_i^{1/2}\right)^{-1} (Y_i - \mu_i) = 0 \qquad (2.7)$$

where $D_i = \partial\mu_i / \partial\hat{\vec{\beta}}^T$, μ_i is the mean function, $V_i = A_i^{1/2} \Re_i(\alpha) A_i^{1/2}$ is a working covariance matrix of the response variable $Y_i = (y_{i1}, y_{i2}, ..., y_{ir})^T$ vector of i=1, 2,…, n individuals observed at the r[th] occasion, $X_i = (X_{i1}, ..., X_{ip})^T$ is a matrix of covariates for individual i, p indicate the number of covariates, $A_i = diag[\text{var}(Y_{i1}), ..., \text{var}(Y_{ir})]$, $\Re_i(\alpha) = corr(Y_i)$ is a working correlation matrix, and α is a vector of parameters associated with a specified model for corr(Y_i). The variance estimators of $\hat{\vec{\beta}}$ can be estimated by the following expression [1, 31, 38]:

$$\sum_{i=1}^{n}\left(D_i^T V_i^{-1} D_i\right)^{-1}\left(\sum_{i=1}^{n} D_i^T (Y_i-\mu_i)(Y_i-\mu_i)^T D_i\right)\left(\sum_{i=1}^{n} D_i^T V_i^{-1} D_i\right)^{-1}.$$

The above estimation is also known as the "sandwich" estimator. The variance of $\hat{\vec{\beta}}$ obtained by the sandwich estimator is consistent [1, 31]. Statistical methods for Gaussian outcome variables are well established, but few statistical methods are available for non-Gaussian outcomes [14].

The GEE approach is widely used to analyze longitudinal survey data, especially data with discrete outcomes [1, 38]. The model-based GEE approach does not take into account the effect of complex survey design (i.e., stratification, clustering, unequal probability of selection, etc.), but it does take into account the intra-class correlation. Liang and Zeger have shown that parameter estimates are asymptotically normal and consistent when the number of clusters increases [42]. Most software has implementations of the GEE approach and the sandwich estimator of variance-covariance matrix of $\hat{\vec{\beta}}$, which makes it a very popular technique for discrete data [43]. The GEE approach has many robust properties for analyzing longitudinal data, but it has some drawbacks when analyzing longitudinal count data [44]. The most commonly used correlation structures are available, such as exchangeable (EXCH) or compound symmetry (CS), first-order auto-regressive (AR(1)), Toeplitz (TOEP), exponential and unstructured (UN) . There is no straightforward way to choose the working correlation structure, even though the GEE approach provides a consistent estimate of the regression parameters when the working correlation structure is misspecified. The working correlation structure, which provides smaller standard errors of parameters, might be the appropriate correlation structure [45]. The log-likelihood ratio test (LRT) can also be used to compare

between two nested correlation structures [38]. A recent study found that the efficiency of the parameter estimates obtained from the GEE approach can be affected by the choice of the working correlation structure [46]. The GEE technique needs to be modified in order to be utilized for statistical analysis of longitudinal complex survey data for the following reasons: stratification, clustering and unequal probability of selection of individual are common features of longitudinal complex survey data, including within-subject repeated measurements. Sampling units may not be independent because of the longitudinal complex survey design. The standard errors, confidence intervals and p-value obtained from standard computer software (SAS, STATA, SPSS) can be invalid because of a lack of independence of within-subject sampling units in longitudinal complex survey data [43, 47]. Rao introduced the quasi-score test for longitudinal survey data using Taylor linearization and jackknife methods, which take into account the complexities of the complex survey design [8].

Let the survey population of size M with S individuals be selected using a stratified multistage sampling design. Let h be the strata (h=1, 2,…., L), k be the cluster (k=1, 2, …., K_h), i denote the individuals and ω_{hki} denote the longitudinal weights to the i^{th} individual in the k^{th} cluster from the h^{th} stratum. The survey independent estimating equations (IEE) of estimators are [7, 48]

$$\hat{U}_{IEE}\left(\vec{\beta}\right)=\sum_{hki\in S_l}\omega_{hki}D_{hki}{}^{T}V_{hki}{}^{-1}(Y_{hki}-\mu_{hki}) \tag{2.8}$$

where S_l denotes the longitudinal sample. The survey GEE estimator proposed by Rao is of the following form [8]:

$$\hat{U}_{GEE}(\vec{\beta})=\sum_{hki\in S_l}\omega_{hki}D_{hki}{}^{T}(\vec{\beta})A_{hki}{}^{-1/2}\left(\vec{\beta}\right)\hat{R}^{-1}A_{hki}{}^{-1/2}\left(\vec{\beta}\right)\left(Y_{hki}-\mu_{hki}\left(\vec{\beta}\right)\right)=0 \tag{2.9}$$

The regression parameter estimators $\hat{\vec{\beta}}_{GEE}$ can be obtained from the survey GEE, which takes into account the effects of complex survey design.

The variance of $\hat{\vec{\beta}}_{GEE}$ can be consistently obtained at $\vec{\beta} = \hat{\vec{\beta}}_{GEE}$ using the following formula:

$$V\left(\hat{\vec{\beta}}_{GEE}\right) = \hat{J}_G^{-1}(\hat{\vec{\beta}}_{GEE}) v\left(\hat{U}_{GEE}\right) J_G^{-1}(\hat{\vec{\beta}}_{GEE}) \tag{2.10}$$

where $\hat{J}_G(\vec{\beta}) = \sum_{hki \in S_l} \omega_{hki} D_{hki}^T(\vec{\beta}) A_{hki}^{-1}(\vec{\beta}) D_{hki}$ at $\vec{\beta} = \hat{\vec{\beta}}_{GEE}$.

The survey GEE approaches have been used in several studies to analyze longitudinal complex survey data [48]. Wald and quasi-score tests were proposed by Rao for longitudinal survey data using Taylor linearization and jackknife resampling methods, which were taken into account because of the nature of complex survey design, including within-subject correlation [8,48]. The formula for estimating variance is

$$V\left(\hat{S}\right) = \sum_h \frac{1}{K_h(K_h - 1)} \sum_k \left(e_{hk}^* - e_{h.}^*\right)\left(e_{hk}^* - e_{h.}^*\right)^T \tag{2.11}$$

where h denotes the h^{th} stratum, k denotes the k^{th} cluster within the h^{th} stratum and

$e_{h.}^* = \sum_k \frac{e_{hk}^*}{K_h}$.

2.3 Use of multilevel modeling–scaled weights

In real-life multistage complex survey data, there is a hierarchical structure of population. For example, individuals are grouped within households, and households are grouped geographically. Repeated measurements of a subject are nested within-subject, and within-subject measurements might be correlated in longitudinal complex survey data. The

dependency between subjects or within subjects or both between and within subject can occur frequently either in cross-sectional or longitudinal complex survey data because of complex survey design. The sampling units are not independent in complex survey data. Traditional statistical methods that are based on the assumption that sampling units are independent are not appropriate for analyzing complex survey data. To analyze large population-based complex survey data (cross-sectional or longitudinal), statistical methods should consider these dependencies between subjects and within subjects in the analysis stage. Observations are assumed to be independent in traditional statistical methods, which might produce biased estimates of parameters in complex survey data analysis [49]. The idea and technique of analyzing multilevel data was first introduced by Mason et al. [107]. To analyze complex survey data, Goldstein proposed multilevel models, which take into account the dependency among individuals as well as the sampling design effects of complex survey data [51].

A number of statistical methods have been developed to analyze complex survey data with hierarchical structures. The multilevel modeling approach is a commonly used statistical method to analyze cross-sectional and longitudinal complex survey data. Goldstein and Raudenbush have made significant contributions to expanding multilevel models for analyzing multistage complex survey data for linear outcomes [51]. Multilevel models are also suitable for discrete outcomes, such as binary and count complex survey data [52]. The term ‘multilevel’ refers to the random variables in the model that vary between units at different levels of the hierarchy [49]. Randomization at the individual level provides more efficient estimates, i.e., smaller standard errors, smaller confidence intervals and more power [53]. Multilevel models are more flexible, and they provide variation between clusters and more efficient parameter estimators compared with traditional techniques [49]. The

individuals may not be independent within clusters in complex survey data, which contradicts the traditional model assumption of independence. Traditional methods (or naïve regression) ignore the dependency between individuals within a cluster [53]. In the multilevel modeling approach, parameter estimation can be biased if the number of level 1 units nested within level 2 is relatively small [54].

In longitudinal complex surveys, there are repeated within-subject measurements, and the data can be unbalanced. The multilevel modeling approach can handle the more realistic missing-at-random (MAR) type, and it might provide unbiased regression coefficients and standard errors for regression coefficients in unbalanced data [53]. Cross-level interaction can also be analyzed by multilevel models. The standard errors of parameter estimators will be smaller and the confidence intervals will be narrower when the dependency between individuals is ignored [13]. Longitudinal data has a hierarchical structure in which repeated measures can be nested within subjects and subjects can be nested within geographical area such as a PSU. Structural equation models (SEMs) based on the multilevel modeling approach can be used to analyze longitudinal complex survey data [9, 55]. Hierarchical linear models, random-intercept or random-coefficient models and variance component models are all known as multilevel models. In multilevel models, the response variable is measured at the lowest level and the explanatory variables can be measured at all levels. Multistage complex survey data arises routinely in different types of fields in which individuals are nested within higher levels. For example, in public health, patients are nested within physicians and physicians are nested within hospitals. Multistage complex survey data may have two or more stages that correspond to the levels of the multilevel models. In order to take into account the unequal probability of selection and non-responses in samples obtained from complex survey data,

probability weights are required to be incorporated in statistical methods. The parameter estimates obtained from complex survey data can be severely biased if the sampling weights are ignored [56]. Weights that are usually available in publicly used complex survey data are not appropriate for multilevel modeling [57]. Unequal probability of selection of sampling units is a common feature of large, complex surveys. Probability weight variables, derived by statistical methodologists, are used to take into account the effect of an unequal probability of selection and the non-response of individuals. It is necessary to have the probability weights for each level of complex survey data in order to use multilevel modeling techniques [6, 25]. A probability-weighted procedure was revealed by Grilli and Pratesi for multilevel binary and ordinal models to reduce the biasing of parameter estimates based on the pseudo maximum likelihood approach [58]. The scaling of weights has a significant influence on parameter estimates and reduces computational problems such as convergence when using multilevel modeling techniques [57, 59]. Several studies have shown that the scaling of weights provides consistently better estimates of parameters, but no gold standard scaling method has been found for the scaling of weights [6, 57, 60]. The scaling of weights is an important tool for decreasing the bias in the estimation of parameters [6, 25]. The ratio between two weights of individuals from different clusters can illustrate oversampling. If the ratio is a meaningful quantity, then scaling might be required [57]. In multilevel modeling based on multilevel pseudo maximum likelihood (MPML), scaling of individual weight levels (level 1) has an influence on parameter estimation but is independent of the scale of level 2 weights [6, 25].

In multilevel data sampling, the units are no longer independent within and between levels. Multistage clustered survey sampling design is used in large health surveys, and the modeling of sampling design is the key issue in estimating the parameters from this type of

sample. Sampling units may not be independent within a cluster. Model-based analyses based on complex survey design provide a biased estimation of parameters [61]. The final sampling or single weights may not approximate the targeted population with the sampled population. Graubard and Korn pointed out that weighted estimation obtained from multistage complex survey data using between-cluster and within-cluster sample weights can be improved [61]. A multilevel modeling technique might be the appropriate approach to incorporate the probability weights for each level of complex surveys.

The standard errors of parameter estimates have a special influence in making valid statistical inference. The impact of cluster sampling on standard errors was investigated by Skinner et al. for longitudinal complex survey data [42]. The findings from the study indicated that the standard errors of regression coefficients can be increased if the impact of the cluster in longitudinal surveys is ignored. The authors also suggested that if the impact of clustering represented by additive random effects in multilevel modeling is used to analyze longitudinal complex survey data, then standard errors can be underestimated. An alternative approach might be to use the GEE to handle the impact of clustering in longitudinal complex surveys. The survey sample selected from a hierarchical population using complex survey design cannot be considered an iid sample because of within-group and between-group correlations between sampling units.

The probability of selection of a sampling unit cannot be equal at different levels of complex survey data. In order to analyze such data using a multilevel modeling approach, the probability weights for sampling units at different levels are required. For example, g is the number of groups (i.e., number of PSUs) selected from G groups and s_g is the collection of n_g sampling units selected from the g^{th} group. Let ω_g be the probability weights for group-

level units and $\omega_{i|g}$ be the conditional sample weights for individuals units. Both probability weights ω_g and $\omega_{i|g}$ are required to analyze complex survey data, whereas only the final probability weights ω_{gi} are commonly available to use for analysis purposes in publicly available complex survey data. Kovačević and Rai described this problem of obtaining the appropriate probability weights for sampling units at different levels and suggested that ω_g is equal to 1 and ω_{gi} is equal to $\omega_{i|g}$ [55]. Asparouhov proposed the multilevel pseudo maximum likelihood (MPML) estimation method for multilevel modeling [57], which is an extension of pseudo maximum likelihood (PML) defined by Skinner MPML is a two-level version of the PML estimator. Missing data can be handled by MPML, with the standard missing-at-random (MAR) assumption. The MPML method produces unbiased parameter estimation, including asymptotic covariance [57]. Several factors have an impact on parameter estimation in multilevel models: cluster sample sizes, informativeness of within-level weights, unequal weighting effects and intra-class correlation (ICC). Kovačević and Rai have shown that if ICC decreases, then biasness of parameter estimates increases [55]. This finding was also supported by Asparouhov [57]. The computational burden increases when the number of random components increases in random-effects or multilevel models.

Weighted estimation using multilevel models is approximately unbiased with larger cluster sizes but severely biased with smaller clusters [57]. The estimation of parameters is more influenced by individual unit levels. This means the probability of inclusion of individuals at each level depends on the response, which may provides the bias estimators of the parameters in standard maximum likelihood estimates [57].

Let y_{ik} be the observed vector of the response variable in cluster k=1, 2,….., K of individual i=1, 2,…., n_k and x_{ik} and x_k be the individual level which indicated level 1 and the cluster level which indicated level 2 covariates, respectively. The level 2 random effect is η_k in cluster k. Let $f(y_{ik} \mid x_{ik}, \eta_k, \theta_1)$ and $\phi(\eta_k \mid x_k, \theta_2)$ denote the density function of y_{ik} and η_k, respectively and $\vec{\theta} = (\theta_1, \theta_2)$ be the vector parameters where θ_1 indicate individual and θ_2 indicate cluster level parameter. The sampling weights for the cluster level and the individual level are $\omega_k = \frac{1}{p_k}$ and $\omega_{i|k} = \frac{1}{p_{i|k}}$, where p_k and $p_{i|k}$ are the probability of selection at the cluster level and the individual level, respectively. The MPML can be defined using the sampling weights of each level as follows:

$$l(\theta_1, \theta_2) = \prod_k \left(\int \left(\prod_i f(y_{ik} \mid x_{ik}, \eta_k, \theta_1)^{\omega_{ik} s_{1k}} \right) \phi(\eta_k \mid x_k, \theta_2) d\eta_k \right)^{\omega_k s_{2k}} \tag{2.12}$$

where s_{1k} and s_{2k} are the scaling constant in level 1 and level 2, respectively [57]. Variance estimators can be obtained by the asymptotic covariance matrix as follows:

$(L''_\omega)^{-1} \left(\sum_j (s_{2k} \omega_k)^2 L'_{\omega j} L'^{\,T}_{\omega j} \right) (L''_\omega)^{-1}$, where ´ and ″ indicate the first and second derivative of the weighted log-likelihood $L_\omega = \log(l)$ [57].

2.4 Monte Carlo Simulation Technique

Simulation techniques are often used to test particular hypotheses, to assess the performance of statistical methods and to identify the true estimation of parameters using computer software [62, 63]. Simulation technique is a numerical method for conducting the experiments based on hypothetical data generated by computer-based software [6, 62, 64].

Researchers commonly perform the simulation studies to (i) assess the properties of parameter estimators, (ii) determine sample sizes, and (iii) test various hypotheses to establish the confidence levels of the results obtained from the analysis of data using statistical methods [62, 65, 66]. Almost half of the articles in the Journal of the American Statistical Association (JASA) used simulation techniques to accomplish their objectives [65]. Simulation studies are widely conducted to assess the performance of a variety of statistical methods in literature [21, 63, 66, 68]. Monte Carlo simulation study is a popular simulation technique that was first studied by De Forest and Stigler (1987), who described Monte Carlo simulation in detail [69]. The usages of the Monte Carlo simulation technique are rapidly expanded because of the widespread availability of computer software. It became an important tool in the development of statistical theory. For example, if the properties of a statistical theory or formula could not be proven analytically, then the Monte Carlo simulation technique would be used to assess the properties of that method.

The third objective of the thesis is to assess the performance of the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique to analyze cross-sectional complex survey data. The RANTBL function in SAS® program is commonly used to generate categorical data and the power of statistical methods are assessed based on the analysis of the generated data [65]. RANBIN and RANPOI are also used to generate categorical data from binomial and poisson distributions respectively in the SAS® program for simulation purposes [67]. Up to date analyzing complex survey data by utilizing two different statistical methods (MM-SW technique and SR-RV estimation technique) does not provide the answer to the question adequately which method performs

relatively better. In order to answer this question the simulated data is generated and analyzed utilizing the both statistical techniques.

2.5 Epidemiology of type 2 diabetes

Type 2 diabetes is a complex chronic condition that occurs when the body does not produce enough insulin or the body cannot properly use the insulin it does produce. There are different types of diabetes, such as type 1, type 2 and gestational. The most common type of diabetes is type 2, which is usually developed among adults. Type 2 diabetes is also known as non–insulin-dependent diabetes. Type 2 diabetes is associated with many life-threatening complications. The most common long-term complications are kidney disease, eye disease, cardiovascular disease (which lead to heart attack and stroke) and diabetic neuropathy (of the feet and lower limbs) [70]. Diabetes is a global epidemic with devastating human, social and economic consequences. It was estimated by the International Diabetes Federation (IDF) in 2007 that 246 million people were suffering from diabetes, and the expected number of diabetic people will be 380 million by 2025 worldwide. According to an IDF report in 2007, the prevalence rate was highest in the Eastern Mediterranean and Middle East Region (9.2%), followed by the North American Region (8.4%). The prevalence of type 2 diabetes was increasing rapidly worldwide. At least 3.8 million deaths occurred directly linked to type 2 diabetes-related causes, including cardiovascular disease. A huge amount of money was spent for treatment of type 2 diabetes globally [70]. Type 2 diabetes was the fifth leading cause of death worldwide [70]. Type 2 diabetes is one of the most important causes of medical expenditures, disability and lost economic growth worldwide.

The etiology of type 2 diabetes is not yet completely known. Studies in India, West Algeria, United Arab Emirates, Qatar, Iran and Brazil reported that the associated risk factors for type 2 diabetes are income, age, smoking status, education, occupation, body mass index, waist circumference, ethnicity and lack of physical activity [15, 71-76]. A study in the USA reported that the prevalence of type 2 diabetes among rural African-American residents was higher than among urban residents [77]. The prevalence of type 2 diabetes is increasing rapidly all over the world as well as in Canada [17, 73]. The health care costs for diabetic people is substantially higher in Canada [78]. In 1996, type 2 diabetes was the cause of death for 5,447 Canadian adults (2,701 males, 2,746 females) [17]. The expected prevalence of type 2 diabetes is 2.4 million by the year 2016 in Canada [16]. Canadian studies have reported that the prevalence of type 2 diabetes was higher among the less educated and in lower earning groups [74]. Studies in Canada have shown that aboriginals are more likely to have type 2 diabetes compared with non-aboriginals. The National Diabetes Surveillance System (NDSS) reported in 2005–2006 that approximately 1.9 million Canadians have type 2 diabetes, with prevalence rate is 5.9%. Among Canadian adults, the death rate was two times higher for those with diabetes than for those without type 2 diabetes, according to an NDSS report. A study in Canada indicated that the age-adjusted mortality rates increased from 12 to 18 deaths per 1000 [17]. Type 2 diabetes is the 7^{th} leading cause of death in Canada [17].
The prevalence of type 2 diabetes varies by country, area (rural/urban), and gender. Limited research has been conducted to determine the prevalence, incidence and trends in type 2 diabetes among the Canadian population. The potential risk factors for type 2 diabetes are sex, age, location of residence (urban/rural), BMI (body mass index), socioeconomic status,

physical activity, education level, etc. Further research is needed to identify the relationship between these risk factors and the prevalence or incidence of type 2 diabetes.

CHAPTER 3

METHODS

Multistage complex surveys are often used to collect data on a large scale to reduce the cost, time and travel of data collection, but this increases the complexity of the statistical analysis [101] . Clustering, stratification, multiple stages of selection, and unequal probability of selection are common features of complex survey design [2]. Statistical methods should take these features into account to obtain valid parameter estimates. Researchers have frequently used standard regression–robust variance (SR-RV) estimation techniques and multilevel modeling–scaled weights (MM-SW) techniques to analyze complex survey data. Both statistical techniques take into account the design effects of complex survey design in order to determine the unbiased parameter estimates. However, the ways in which these two statistical methods take into account these design effects at the analysis stage are different. A few studies have been conducted to determine the advantages and disadvantages of standard regression–robust variance estimation technique and multilevel modeling–scaled weights technique in order to analyze multistage complex survey data, but no definite conclusions have been drawn [10, 12].

The primary goal of this thesis was to compare the SR-RV estimation technique with the MM-SW technique after taking into account the common features of complex survey design, including weight adjustments. In this chapter, the standard regression–robust variance estimation technique (i.e., Taylor linearization and bootstrapping) and the multilevel modeling–scaled weights technique are discussed in detail.

3.1 Statistical methods to accomplish Objective 1

To explore the usage of the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique to analyze cross-sectional complex survey data.

3.1.1 Standard regression for cross-sectional complex survey data

Let $Y = (y_1, y_2, ..., y_n)$ be the vector the response variable and $\vec{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ip})$ denotes the covariates for the i^{th} individual where i=1,2,3,…, n,. Let the response variable of interest y_i be dichotomous (0 or 1) where '0' represents 'has no disease' and '1' represents 'has disease', the probability of y_i having a value of 1 is μ_i.

The logistic regression models with n data points can be written as

$$\text{Logit } [\Pr(y_i = 1 | \vec{x}_i)] = \log\left(\frac{\mu_i}{1-\mu_i}\right) = \vec{x}_i\vec{\beta} \quad (3.1)$$

where $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ is the vector of the regression coefficients.

3.1.1.1 Parameter Estimation

The maximum likelihood estimation (MLE) technique is used to estimate the regression parameters. The log likelihood function for a binary outcome can be written as

$$L(\vec{\beta}) = \sum_{i \in S} y_i \ln F(\vec{x}_i\vec{\beta}) + \sum_{i \in S} (1-y_i) \ln\{1 - F(\vec{x}_i\vec{\beta})\} \quad (3.2)$$

where $F(\vec{x}_i\vec{\beta}) = \dfrac{\exp(\vec{x}_i\vec{\beta})}{1+\exp(\vec{x}_i\vec{\beta})}$ and S is the set of all observed individuals.

The vector of the regression coefficients $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ can be estimated from the p likelihood equations, which are obtained by differentiating the log likelihood function (3.2) with respect to the regression coefficients. The set of score equations are as follows:

$$\frac{\partial L(\vec{\beta})}{\partial \vec{\beta}} = 0 \tag{3.3}$$

The regression coefficients and their variance and covariance estimates can be obtained from the score equations $\frac{\partial L(\vec{\beta})}{\partial \vec{\beta}} = 0$ and covariance matrix $\sum = \mathrm{I}^{-1}(\vec{\beta})$, where

$$\mathrm{I}\ (\vec{\beta}) = -E\left(\frac{\partial^2 L\ (\vec{\beta})}{\partial \vec{\beta}^2} \right).$$

The pseudo likelihood function is a special type of likelihood function. The sampling weights for sample elements are required to construct the pseudo likelihood function. Let ω_i be the sampling weights for the i[th] individual. The log pseudo likelihood function can be written as

$$L_w(\vec{\beta}) = \sum_{i \in S} \omega_i y_i \ln F(\vec{x}_i \vec{\beta}) + \sum_{i \in S} \omega_i (1 - y_i) \ln \{1 - F(\vec{x}_i \vec{\beta})\} \tag{3.4}$$

where $F(\vec{x}_i \vec{\beta}) = \frac{\exp(\vec{x}_i \vec{\beta})}{1 + \exp(\vec{x}_i \vec{\beta})}$. The vector of the regression coefficients $\vec{\beta} = (\beta_0, \beta_1, ..., \beta_p)$ can be estimated from the p likelihood equations or the score equations, which are obtained by differentiating the log pseudo likelihood function with respect to the regression coefficients. The set of score equations are as follows:

$$\frac{\partial L_\omega(\vec{\beta})}{\partial \vec{\beta}} = 0 \tag{3.5}$$

The regression coefficients and their variance and covariance estimates can be obtained from the score equations $\frac{\partial L_w(\vec{\beta})}{\partial \beta} = 0$ and the covariance matrix $Cov(\vec{\beta}) = \mathrm{I}_w^{-1}(\vec{\beta})$, where $\mathrm{I}_w(\vec{\beta}) = -E\left(\frac{\partial^2 L_w(\vec{\beta})}{\partial \vec{\beta}^2}\right)$.

3.1.1.2 Variance-covariance estimation

It is essential to estimate the correct standard errors to make valid inferences because they play a key role in testing the null hypotheses. Ignoring design effects may lead to underestimation of the standard errors of parameter estimates and consequently to the inaccurate rejection of the null hypotheses.

The standard errors will be large and the confidence intervals will be wide if the effects of stratification are ignored when analyzing complex survey data [1, 2]. The standard errors will be small and the obtained results will often be significant if the effects of clustering are ignored in multistage complex survey data [2]. Taylor linearization and resampling methods (i.e., jackknifing, balanced repeated replications (BRR) and Rao-Wu bootstrapping) are commonly used to estimate the variance of parameter estimators, which are discussed in detail in sections 3.1.1.2.1 and 3.1.1.2.2.

3.1.1.2.1 Taylor linearization

Let $L_\omega(\vec{\beta})$ be a smooth function of $\vec{\beta}=(\beta_0,\beta_1,\beta_2,...,\beta_p)$, and $\hat{\vec{\beta}}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_p)$ be an estimator vector of $\vec{\beta}=(\beta_0,\beta_1,\beta_2,...,\beta_p)$ where $L_\omega(\vec{\beta})$ is a log pseudo likelihood function. The variance can be obtained by linearization using a Taylor expansion of $\hat{G}(\vec{\beta})$ at $\vec{\beta}=\hat{\vec{\beta}}$, where $\hat{G}(\vec{\beta})=\frac{\partial L_\omega}{\partial\vec{\beta}}$. Binder presented the theory of variance estimation using Taylor expansion for complex survey design [30]. Let $\vec{\beta}$ be the solutions of the set of estimating equations $\hat{G}(\vec{\beta})=0$. The variance of $\vec{\beta}$ can be obtained by Taylor expansion of $\hat{G}(\vec{\beta})$ at $\vec{\beta}=\hat{\vec{\beta}}$, where $\vec{\beta}$ is the regression parameter vector value.

$$0=\hat{G}(\vec{\beta})\cong\hat{G}\left(\hat{\vec{\beta}}\right)+\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)}{\partial\hat{\vec{\beta}}}\left(\vec{\beta}-\hat{\vec{\beta}}\right)+\frac{1}{2}\frac{\partial^2\hat{G}}{\partial\vec{\beta}^2}\left(\vec{\beta}-\hat{\vec{\beta}}\right)^2$$

$$\hat{G}\left(\hat{\vec{\beta}}\right)\cong-\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)}{\partial\hat{\vec{\beta}}}\left(\vec{\beta}-\hat{\vec{\beta}}\right) \tag{3.6}$$

Taking the variance both sides of (3.6), we obtain

$$V\left(\hat{\vec{\beta}}\right)=\left\{\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)}{\partial\hat{\vec{\beta}}}\right\}V(\vec{\beta})\left\{\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)}{\partial\hat{\vec{\beta}}}\right\}^T$$

$$\hat{V}(\vec{\beta})=\left[\left\{\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)}{\partial\hat{\vec{\beta}}}\right\}^{-1}V\left(\hat{\vec{\beta}}\right)\left\{\frac{\partial\hat{G}\left(\hat{\vec{\beta}}\right)^T}{\partial\hat{\vec{\beta}}}\right\}^{-1}\right] \tag{3.7}$$

Stratum and PSU identifiers are not commonly available in publicly used data files for confidentially reasons; these identifiers are required for variance estimation.

3.1.1.2.2 Bootstrap variance estimation

Bootstrapping is a re-sampling technique that produces artificial simple random sampling data from observed data with the same sample size [33]. Rao, Wu and Yue [28] extended the bootstrap procedure for multistage complex surveys; their approach can take into account the design features (i.e., stratification, clustering). The bootstrap variance estimation procedure derived from the following steps is used to determine the variance of parameter estimates for multistage complex survey data.

Step 1: Let the total number of bootstrap independent samples from the observed sample be B (for example, B=500) and the bootstrap weights can be calculated for each sampled unit with the replacement of K_h-1 clusters from K_h sampled clusters for each stratum by

$$\omega^*_{hki} = \frac{K_h}{K_h - 1} m^*_{hk} \omega_{hki} \tag{3.8}$$

where m^*_{hk} is the number of times the $(hk)^{th}$ cluster appears.

Step 2: Replace the bootstrap weights $\omega_{hki}(b)$ with the sampling weights in the estimating equations or the score equations and calculate the bootstrap estimate $\hat{\tilde{\beta}}^*_b$ where b=1,2,...,B.

Step 3: Repeat step1 and step2 B times, and calculate the bootstrap estimates $\hat{\beta}_{(1)}{}^*, \hat{\beta}_{(2)}{}^*, ..., \hat{\beta}_{(B)}{}^*$.

Step 4: Obtain the bootstrap variance estimators for $\hat{\tilde{\beta}}$ [7] with the following equation:

$$\text{var}\left(\hat{\theta}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}_{(b)}{}^{*} - \hat{\theta}_{(\cdot)}{}^{*}\right)^{2} \text{ where } \hat{\theta}_{(\cdot)}{}^{*} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}_{(b)}{}^{*} \tag{3.9}$$

3.1.2 Multilevel models for cross-sectional complex survey data

Multilevel models are often used to analyze multistage complex survey data when clustering, stratification, and unequal probability of selection are involved. The responses can be correlated because of the unobserved heterogeneity between clusters, which should be taken into account in order to make valid statistical inferences. In multistage complex survey data, the sample units may not be independent within a cluster or between clusters. In this thesis, the response variable is binary. The logistic random-intercept model or the logistic random-coefficient model can be used to analyze complex survey data based on multilevel pseudo maximum likelihood (MPML). We may require the sample weight for each level to analyze multistage complex survey data. The unequal probability of selection is taken into account using the sampling weights of individuals at each level of multistage complex survey data. Parameter estimates obtained from the analysis of complex survey data can be severely biased if the sampling weights are ignored [56].

Let us consider the binary response variable y_{ik}, which was measured at the lowest level in hierarchical data structures and $\vec{x}_{ik}$, which is the explanatory variable on the i^{th} unit in level 1 within the k^{th} unit in level 2. Let $\vec{x}_{ik}$ be the vector of covariates, and let $\vec{\beta}$ be the vector of fixed regression coefficients. Let $\zeta_k^{(2))}$ be the random effects varying over clusters k, where $\zeta_k^{(2)} \sim \text{N}(0,\psi)$. A two-level generalized linear mixed model with linear predictors can be defined as

$$\text{Logit}\left[pr\left(y_{ik}=1 \mid \vec{x}_{ik}, \zeta_k^{(2)}\right)\right] = \vec{x}_{ik}\vec{\beta} + \vec{x}_{ik}\zeta_k^{(2)} \quad (3.10)$$

The following two types of multilevel models are commonly used for binary outcomes:

i) Multilevel logistic random-intercept models, discussed in section 3.1.2.1.

ii) Multilevel logistic random-coefficient models, discussed in section 3.1.2.2.

3.1.2.1 Multilevel logistic random-intercept models

The models in which the overall level of response is considered to vary over clusters after adjusting for potential covariates are known as multilevel random-intercept models. Consider the multilevel (two-level) logistic random-intercept model for unit i (level 1) within the cluster k (level 2). For example, in the Canadian Heart Health Survey (CHHS), level 1 is an individual and level 2 is a primary sampling unit (PSU), also known as a cluster. The multilevel logistic random-intercept model is [25]

$$\begin{aligned}
&\text{Logit}\left[\Pr\left(y_{ik}=1 \mid \vec{x}_{ik}, \zeta_k^{(2)}\right)\right] \\
&= \nu_{ik} \\
&= \vec{x}_{ik}\vec{\beta} + \zeta_k^{(2)} \\
&= \beta_0 + \beta_1 x_{1ik} + \beta_2 x_{2ik} + \ldots + \beta_p x_{pik} + \zeta_k^{(2)} \\
&= \left(\beta_0 + \zeta_k^{(2)}\right) + \beta_1 x_{1ik} + \beta_2 x_{2ik} + \ldots + \beta_p x_{pik} \quad (3.11)
\end{aligned}$$

where $\zeta_k^{(2)}$ are the random intercepts and are considered random variables. The effects of unobserved heterogeneity can be represented by the random parameters $\zeta_k^{(2)}$, with $\zeta_k^{(2)} \mid x_{ik} \sim N(0, \psi)$. The random intercepts $\zeta_k^{(2)}$ are independent across the level 2 units. The random-intercept model is assumed to capture the combined effects of the fixed effects $\vec{\beta}$ and the

random effects $\zeta_k^{(2)}$. The random-intercept models are parallel to each other because of the constant slope $\vec{\beta}$ for every model.

3.1.2.2 Multilevel logistic random-coefficient models

The models where the overall level of response and the effects of covariates are considered to vary over clusters after controlling for covariates are known as multilevel random-coefficient models. Consider the multilevel (two-level) logistic random-coefficient model for unit i (level 1) within the cluster k (level 2). As mentioned above, in the CHHS, level 1 corresponds to an individual and level 2 corresponds to a PSU. The multilevel logistic random-coefficient model is as follows [25]:

$$\begin{aligned}
&\text{Logit}\left[\Pr\left(y_{ik}=1 \mid \vec{x}_{ik}, \zeta_k^{(2)}\right)\right] \\
&= \nu_{ik} \\
&= \vec{x}_{ik}(\vec{\beta}+\zeta_k^{(2)}) \\
&= \beta_0 + \beta_1 x_{1ik} + \beta_2 x_{2ik} + \ldots + \beta_p x_{pik} + \zeta_k^{(2)} + \zeta_k^{(2)} x_{1ik} + \ldots + \zeta_k^{(2)} x_{pik} \\
&= \left(\beta_0 + \zeta_k^{(2)}\right) + \left(\beta_1 + \zeta_k^{(2)}\right) x_{1ik} + \left(\beta_2 + \zeta_k^{(2)}\right) x_{2ik} + \ldots + \left(\beta_p + \zeta_k^{(2)}\right) x_{pik}
\end{aligned} \tag{3.12}$$

where $\vec{x}_{ik}$ are uncorrelated with $\zeta_k^{(2)}$ and $\zeta_k^{(2)}$ are independent across level 2 units (k).The term $\zeta_k^{(2)}\vec{x}_{ik}$ indicates the interaction between the clusters and the covariates. The random intercept and the random slope have a bivariate normal distribution with a mean of zero and a covariance matrix ψ.

3.1.2.3 Parameter Estimation for the Multilevel Model

3.1.2.3.1 Maximum Likelihood Estimation (MLE)

Let $\vec{\beta}$ be the vector of regression parameters. The usual marginal maximum log-likelihood function can be written as

$$L(\vec{\beta}) = \sum_{k=1}^{K} \log \int_{-\infty}^{\infty} \exp\left\{ \sum_{i=1}^{n_{kj}^{(1)}} \log f(y_{ik} \mid \varsigma_k^{(2)}) \right\} g(\varsigma_k^{(2)}) d\varsigma_k^{(2)} \tag{3.13}$$

where $\varsigma_k^{(2)} \sim N(0, \psi)$ and $g(\varsigma_k^{(2)})$ is the normal density function.

In multistage complex surveys, the probabilities of selection of units at the corresponding levels are unequal. The usual maximum log-likelihood estimates are biased without taking into account the unequal probability of selection [6]. The pseudo maximum log-likelihood algorithm can accommodate the probability weights and reduce the bias of parameter estimates.

3.1.2.3.2 Multilevel Pseudo Maximum Likelihood (MPML)

Let us consider the two-stage sampling design in which π_k (k= 1, 2, …,K) is the probability of selection of a level 2 unit and $\pi_{i|k}$ (i=1, 2, ….., $n^{(1)}$) is the probability of selection of the i^{th} unit in level 1 within the k^{th} cluster in level 2. Let $\omega_k = 1/\pi_k$ and $\omega_{i|k} = 1/\pi_{i|k}$ be the inverse probability of selection of the k^{th} unit in level 2 and the i^{th} unit in level 1 within the k^{th} unit in level 2, respectively. The multilevel pseudo log-likelihood can be defined [6] as

$$L(\vec{\beta}) = \sum_{k=1}^{K} \omega_k \log \int_{-\infty}^{\infty} \exp\left\{ \sum_{i=1}^{n_k^{(1)}} \omega_{i|k} \log f(y_{ik} \mid \varsigma_k^{(2)}) \right\} g(\varsigma_k^{(2)}) d\varsigma_k^{(2)} \tag{3.14}$$

where $g(\varsigma_k^{(2)})$ is the normal density function, $\varsigma_k^{(2)} \sim N(0,\psi)$ and $\vec{\beta}$ is the vector of parameters.

The probability weights of units for each level are incorporated in the above multilevel pseudo maximum log-likelihood algorithm. The level 1 weight can be varied between elementary units, and the parameter estimates can be biased. The scaling of level 1 weight has an effect on the estimates of the regression coefficients and their variances, especially when the responses are binary [57]. The likelihood function which is the joint probability of responses with given all potential covariates does not have a closed form in generalized linear mixed models, and approximate methods are required to evaluate it. A procedure is described in the next section.

3.1.2.3.3 Adaptive Quadrature

The likelihood function generally does not have a closed form in generalized linear mixed models. It is often complicated to estimate parameters from the likelihood function because of the intractable integral. The Gauss–Hermite quadrature approach is commonly used to maximize the likelihood function. In random-effects models, the computational burden increases when the number of random components increases [79]. This technique provides biased estimates with large cluster sizes [25]. The alternative of Gauss–Hermite quadrature is the adaptive quadrature approach, which consists of scaling and translating the quadrature locations.

Let $Y = (y_{1k}, y_{2k}, ..., y_{n_k k})$ be the vector of response and $\vec{x}_k$ be the vector of covariates. The likelihood function—the joint probability of all responses, given the covariates—is

$$L(\vec{\beta}) = \prod_{k=1}^{K} pr(y_{1k}, y_{2k}, ..., y_{n_k k} \mid \vec{x}_k) \tag{3.15}$$

where

$$pr(y_{1k}, y_{2k,}, ..., y_{n_k k} \mid \vec{x}_k) = \int pr(y_{1k,} y_{2k}, ..., y_{n_k k} \mid \vec{x}_k, \varsigma_k^{(2)}) g(\varsigma_k^{(2)}; 0, \psi) d\varsigma_k^{(2)} \tag{3.16}$$

and $pr(y_{1k}, y_{2k}, ..., y_{n_k k} \mid \vec{x}_k, \zeta_k^{(2)}) = \prod_{i=1}^{n_k} pr(y_{ik} \mid \vec{x}_k, \zeta_k^{(2)}) = \prod_{i=1}^{n_k} \frac{\exp(\vec{\beta}\vec{x}_k + \zeta_k^{(2)})^{y_{ik}}}{1 + \exp(\vec{\beta}\vec{x}_k + \zeta_k^{(2)})}$

The normal density function of $\varsigma_k^{(2)}$ is $g(\varsigma_k^{(2)}, 0, \psi)$, with a mean zero and variance ψ.

The right hand side of Equation (3.15) can be approximated by a sum of R terms with e_r (r = 1, 2,…, R) for $\varsigma_k^{(2)}$ and by replacing ω_r with $g(\varsigma_k^{(2)}, 0, \psi)$:

$$pr(y_{1j}, y_{2k,}, ..., y_{n_k k} \mid \vec{x}_k) = \sum_{r=1}^{R} pr(y_{1k}, y_{2k}, ..., y_{n_k k} \mid \vec{x}_k, \varsigma_k^{(2)} = e_r) \omega_r$$

where e_r and ω_r are the Gauss–Hermite quadrature locations and the weights respectively. The locations are rescaled and translated as $e_{rq} = a_q + b_q e_r$, where a_q and b_q are cluster-specific constants. These transformations go along with the weights ω_r, which also depend on a_q and b_q. The adaptive quadrature approximate approach uses the GLLAMM procedure in STATA. The probability weights of units for each level are incorporated in the above multilevel pseudo-likelihood algorithm. The level 1 weights can vary between elementary units, and the parameter estimates can be biased. The scaling of level 1 weights has an effect on the estimates of the regression coefficients and their variances, especially when the responses are binary [57].

3.1.2.3.4 Relationship between regression coefficients obtained from multilevel modeling and standard regression

The relationship between the regression coefficients obtained from multilevel models (random-effects models) and standard regression can be defined as $\beta_{sr} \approx \frac{\beta_{ml}}{\sqrt{1+0.3\sigma^2}}$, β_{sr} is estimated using a marginal model (standard regression) , β_{ml} is estimated using a random-effects model (multilevel modeling) and σ^2 is estimated between-subject variation [38]. The above relationship indicates that regression coefficient estimates can be higher in the multilevel modeling–scaled weights technique than in the standard regression–robust variance estimation technique. If the between-subject variation (σ^2) is higher, then the regression coefficient estimates (β_{ml}) obtained from the multilevel modeling–scaled weights technique will be almost always higher than the regression coefficient estimates (β_{sr}) obtained from the standard regression–robust variance estimation technique.

3.1.2.3.5 Scaling of weights

The purpose of scaling is to reduce the bias of parameter estimators [57]. There are several types of scaling methods available for the scaling of level 1 weights [6].

Method 1:

Let $\omega_{i|k}$ be the level 1 weights. The scale factors are defined as $a_1^{(1)} = \frac{\sum_{i=1}^{n^{(1)}} \omega_{i|k}}{\sum_{i=1}^{n^1_k} (\omega_{i|k})^2}$. The scaling of weights is equal to $\omega_{.|k} = \sum_{i=1}^{nk} a_1^{(1)} \omega_{i|k}$,

where $i = 1,2,...,n^{(1)}$ denotes level 1 units and $k = 1,2,...,K$ denote the level 2 units.

Method 2:

The scaling factor $a^1 = n_k^{(1)} / \omega_{\cdot|k}$ sets the apparent cluster size $\omega_{\cdot|k}{}^a$ equal to the actual cluster size $n_k^{(1)}$.

Method 3:

The new level 2 weights are created as $\omega_k^* = \sum_{i=1}^{n_j^{(1)}} \omega_{i|k}\omega_k$, and the level 1 weights are $\omega_{i|k}^* = 1$.

3.1.2.4 Variance estimation

3.1.2.4.1 Sandwich estimator of the standard errors

The form of the covariance matrix can be defined [6] as

$$\mathrm{cov}(\vec{\beta}) = I^{-1} J I^{-1} \tag{3.17}$$

where I is the Fisher information matrix and $J \equiv E\left\{ \frac{\partial L_\omega(y;\vec{\beta})}{\partial \vec{\beta}} \frac{\partial L_\omega(y;\vec{\beta})}{\partial \vec{\beta}^T} \right\}_{\vec{\beta}=\hat{\vec{\beta}}}$,

where $\vec{\beta}$ is the vector of the parameters and the pseudo log-likelihood function is

$$L_\omega(y;\vec{\beta}) = \sum_{k=1}^{n^{(2)}} \omega_k \log \int_{-\infty}^{+\infty} \exp\left[\sum_{i=1}^{n^{(1)}} \omega_{i|k} \log f(y_{ik} \mid \zeta_k) \right] g(\zeta_k) d\zeta_k$$

The gradient of the pseudo log-likelihood function is the sum of the independent clusters: $\frac{\partial L_\omega(y;\vec{\beta})}{\partial \vec{\beta}} = \sum_{t=1}^{n^{(q)}} \omega_t{}^{(q)} \frac{\partial L_\omega{}^{(q)}(y\ ;\vec{\beta})}{\partial \vec{\beta}} \equiv \sum_{t=1}^{n^{(q)}} S_t(\vec{\beta})$

where q denote the level.

Let S_{hkt} be the weighted score vector of the top-level unit t in stratum $h(h=1,2,...,H)$ and cluster $k(k=1,2,...,K_h)$ where $t=1,2,...,N_{hk}$ individuals within stratum h and cluster k. The gradient of the pseudo log-likelihood can be written as

$$\frac{\partial L_\omega(y;\vec{\beta})}{\partial \vec{\beta}}\Big|_{\vec{\beta}=\hat{\vec{\beta}}} = \sum_{h=1}^{H}\sum_{k=1}^{K_h}\sum_{t=1}^{N_{hk}} \omega_{ght}\frac{\partial L_\omega\left(y;\vec{\beta}\right)}{\partial \vec{\beta}} \equiv \sum_{h=1}^{H}\sum_{k=1}^{K_h}\sum_{t=1}^{N_{hk}} S_{hkt}\ .$$

After taking stratification and clustering into account, the covariance matrix will be [6]

$$J = \sum_{h=1}^{H}\frac{K_h}{K_h-1}\sum_{k=1}^{K_h}\left(S_{hk.}-\bar{S}_{h..}\right)\left(S_{hk.}-\bar{S}_{h..}\right)^{/} \tag{3.18}$$

where $S_{hk.}=\sum_{t=1}^{N_{hk}} S_{hkt}$, $\bar{S}_{h..}=\frac{1}{K_h}\sum_{k=1}^{G_h} S_{hk.}$

The statistical methods used to analyze the data based on complex surveys are shown in the flow chart in figure 3.1.

3.1.3 Goodness-of-fit test for logistic regression

A logistic regression model is used to determine the probability of an event (type 2 diabetes) for a dichotomous (yes, no) outcome as a function of the covariates. It is necessary to investigate how well the predicted logistic regression model fits the data after fitting a logistic regression model. The goals of the goodness-of-fit test are to see whether the model fits the observed data adequately and to describe the association between the outcome and the potential risk factors. The goodness-of-fit that is measured based on residuals tests the overall differences between the observed and fitted values. The small differences between the observed and fitted values indicate that the model fits observed data adequately. Any conclusions or results obtained from the regression analysis might be incorrect or misleading

if the goodness-of-fit test for the estimated model is not performed. Several methods are available to assess the goodness-of-fit test for the predicted logistic regression model. The Hosmer–Lemeshow goodness-of-fit test is commonly used to assess the model fit, but it is not implemented in all publicly used statistical software programs for survey data. The F-adjusted mean residual test is implemented in STATA for survey data. The F-adjusted mean residual test procedures are discussed below.

3.1.3.1 Goodness-of-fit test for survey sample

Let y_{ik} be the observed outcome for the i[th] individual within the k[th] primary sample unit (PSU) and $\hat{\pi}(x_{ik})$ be the predicted values. The residuals $\hat{r}_{ik} = y_{ik} - \hat{\pi}(x_{ik})$, which are the differences between the observed and predicted values, will indicate the lack of fit. Small differences between the observed and predicted values indicate a better fit. In this approach, the observations are grouped into deciles based on their estimated probabilities. Let $\hat{M} = (\hat{M}_1, \hat{M}_2, ..., \hat{M}_{10})$ be the vector of estimates of the mean residuals, where

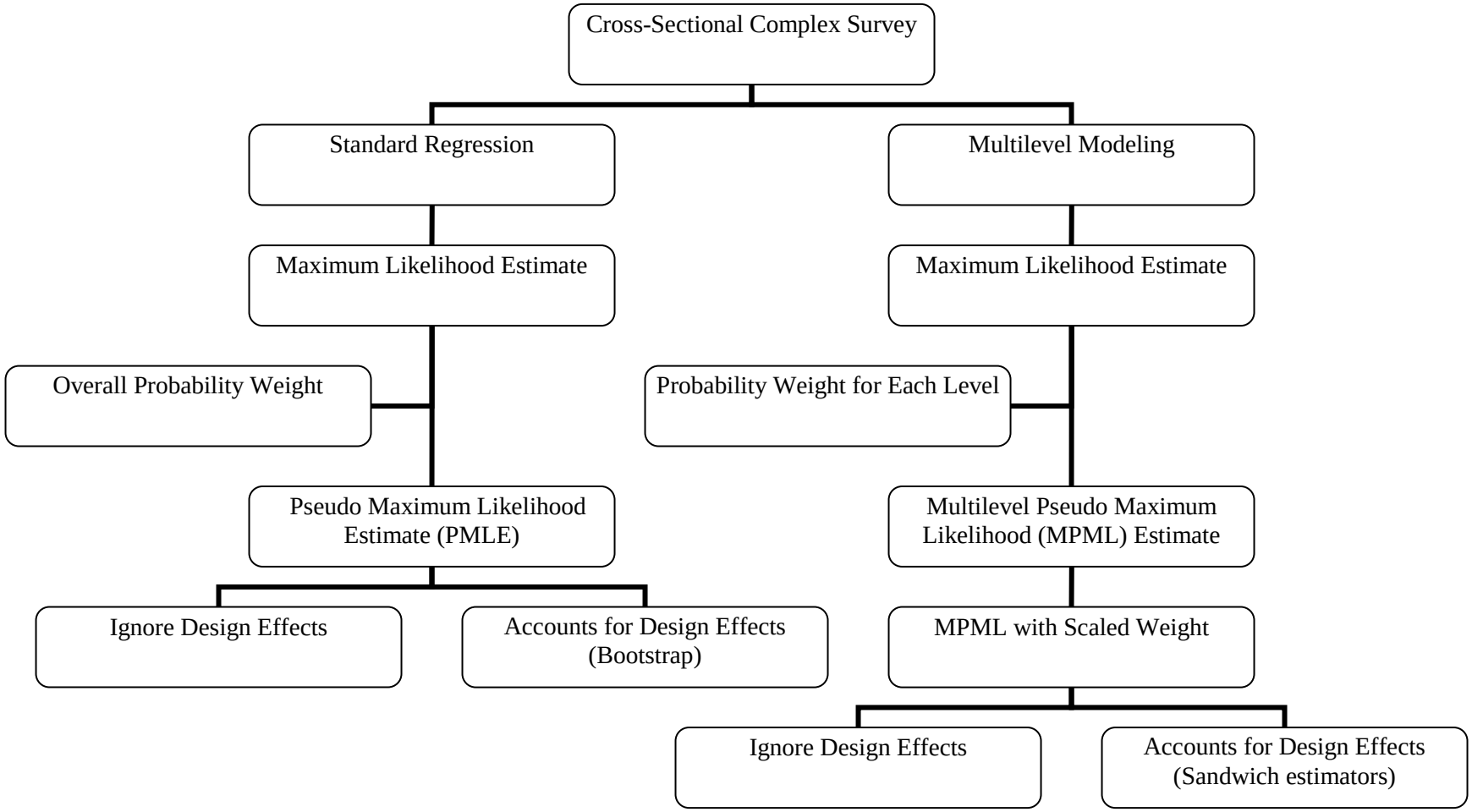
$\hat{M}_1 = \sum_k \sum_i \omega_{ik} \hat{r}_{ik} / \sum_k \sum_i \omega_{ik}$ for the smallest 10% of the $\hat{r}_{ik}$ values,

$\hat{M}_2 = \sum_k \sum_i \omega_{ik} \hat{r}_{ik} / \sum_k \sum_i \omega_{ik}$ for the second smallest 10% of the $\hat{r}_{ik}$ values, and

$\hat{M}_{10} = \sum_k \sum_i \omega_{ik} \hat{r}_{ik} / \sum_i \sum_k \omega_{ik} k$ for the largest 10% of the $\hat{r}_{ik}$ values. Here, ω_{ik} represents the sampling weights for the indicated deciles of risk. The F-corrected Wald statistic can be defined as $F = \frac{(f - g + 2)}{(fg)} W$, which is approximately F-distributed with $g - 1$ numerator degrees of freedom and $f - g + 2$ denominator degrees of freedom [80]. Here, f represents

the number of sampled clusters minus the number of strata, g (here g=10) represents the number of categories, and $\hat{W} = \hat{M}^T \left\{ \hat{V}(\hat{M})^{-1}_{g \times g} \right\} (\hat{M})$, which is also known as the Wald test statistic [80]. The variance-covariance matrix $\hat{V}\left(\hat{M}\right)_{g \times g}$ can be obtained based on first-order Taylor series approximation.

Figure 3.1 Flow chart of statistical methods used to accomplish Objective 1



3.2 Statistical Methods to accomplish Objective 2

3.2.1 Standard regression for longitudinal complex survey data

Longitudinal studies are great resources for understanding the development of disease among individuals. Genetic, environmental, social, and behavioral factors are common sources of heterogeneity among individuals who develop diseases. Longitudinal studies can help determine the change in health outcomes among individuals over time. The primary goal of longitudinal studies is to determine the longitudinal changes over time in the outcome variables of interest and the associated risk factors. Repeated measurements of responses for the same individuals over time are a special feature of longitudinal studies. The repeated measurements of responses for the same individual can be correlated. Between-individual heterogeneity, within-individual biological variation, and measurement errors can be sources of variability that contribute to correlations between pairs of response measurements for the same individual. Statistical methods should take into account the within-individual correlations among repeated measurements of responses, including the effect of complex survey design (i.e., stratification, clustering and unequal probability of selection) to obtain valid parameter estimates. There are several statistical methods available to analyze longitudinal complex survey data. The MM-SW technique and the SR-RV estimation technique are commonly used to analyze longitudinal complex survey data. Both statistical methods take into account the effect of complex survey design, including within-subject correlations.
The second objective was to compare the MM-SW technique and the standard regression–robust variance estimation technique to analyze longitudinal complex survey data. In this section, we have discussed the MM-SW technique and the SR-RV estimation technique in detail.

3.2.1.1. Marginal models for binary outcome

The special characteristic of a longitudinal study is the repeated measurement of responses for the same individual over time, whereas a single measurement is taken per individual in a cross-sectional study. The repeated measurements of responses for the same individual are not independent, which should be taken into account when fitting longitudinal data. Marginal models are one of the standard regression techniques used to fit longitudinal complex survey data.

Marginal models, an extension of generalized linear models (GLM), are commonly used to analyze longitudinal data. Marginal models determine the mean response, depending on the covariates of interest, and are also known as population-averaged (PA) models. The advantage of marginal models is that no distributional assumptions are required for the vector of responses [1]. The usual likelihood function is not useful for estimating parameters because of the need to avoid distributional assumptions in the vector of responses. Generalized estimating equations (GEE) based on the quasi-likelihood function are an alternative to the usual likelihood equations and can be used to estimate the regression parameters without knowing the distribution of the response vector [38, 81].

In longitudinal data, the vector of response variable and the vector of covariates for each individual are defined as: $Y_i = (Y_{i1},....,Y_{ir})^T$ (i=1, 2,…, n) and $X_{ir} = (x_{ir1}, x_{ir2},..., x_{irp})^T =$

$$\begin{pmatrix} x_{i11} & x_{i12} & \dots & x_{i1p} \\ x_{i21} & x_{i22} & \dots & x_{i2p} \\ \dots & \dots & \dots & \dots \\ x_{iR1} & x_{iR2} & \dots & x_{iRp} \end{pmatrix}_{R\times P}$$ (i=1, 2,…, n; r=1, 2,…, R), respectively,

where r=1,2,…,R denotes the number of repeated measurements within individual i=1,2,…,n.

The response variable of interest Y_i could be continuous, binary or count. Let n be the observed number of individuals repeatedly measured over time. The main purpose of marginal models is to make inferences about the population mean of the response vector as conditioned by the vector of the covariates and the within-individual correlation from repeated measurements. The general specification of a marginal model for longitudinal data is that the conditional expectation of each response $E(Y_{ir} \mid X_{ir}) = \mu_{ir}$ can be connected with the vector of covariates by using appropriate link function $g(\mu_{ir}) = \eta_{ir} = X^T{}_{ir}\vec{\beta}$, where $\vec{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is the vector of regression coefficients. The variances of each response Y_{ir}, given covariates X_{ir}, are defined as $Var(Y_{ir} \mid X_{ir}) = \phi v(\mu_{ir})$, where ϕ is known as the scale parameter, which may be known or need to be determined, and the variance function $v(\mu_{ir})$ depends on the mean responses. The last component of marginal models is the within-subject association due to the repeated measurement of responses from the same individual, and it can be determined by the covariance matrix.

In this thesis, our response variable of interest Y_{ir} is binary (0, 1), where '0' represents 'failure' and '1' represents 'success'. The probability of Y_{ir} with value '1' (success) is μ_{ir}, i.e., $E(\mathrm{Y}_{ir}) = \Pr(\mathrm{Y}_{ir} = 1) = \mu_{ir}$.

The marginal models are specified by the following logistic regression models:

$$\text{logit}[\Pr(\mathrm{Y}_{ir} = 1)] = \log\left(\frac{\mu_{ir}}{1-\mu_{ir}}\right) = \vec{\beta}_s{}^T X_{is} + \vec{\beta}_t{}^T X_{it} \qquad (3.19)$$

where r=1, 2,…, R (occasions) and i=1, 2,…., n (individuals), $\vec{\beta}_s{}^T$ is a vector of stationary covariates, $\vec{\beta}_t{}^T$ is a vector of time-varying covariates, X_{is} is a design matrix of stationary covariates and X_{it} is a design matrix of time-varying covariates. The variance of the response

depends on the mean response, i.e., $Var(Y_{ir}) = \mu_{ir}(1-\mu_{ir})$. The within-subject association can be defined by an appropriate covariance structure.

3.2.1.1.1 Generalized Estimating Equations (GEE)

In marginal models, a distributional assumption for the response variable of interest is not required. An alternative estimating equation for usual likelihood equations is required because of the avoidance of the distributional assumption of the response variable. Generalized estimating equations (GEE) are the alternative estimating equations in marginal models for estimating parameters when analyzing longitudinal data [81].

The generalization and extension of the usual likelihood function for univariate responses by incorporating the covariance matrix of the vector of responses for longitudinal data is the main reflection of GEE for generalized linear models (GLM). The association among the repeated measurements depends on the mean response (μ_{ir}) and the correlations between pairs of responses for the same individual. The covariance matrix can be defined as $V_i = A_i^{\frac{1}{2}} Corr(Y_i) A_i^{\frac{1}{2}}$, where A_i is a diagonal matrix with $Var(Y_i) = v(\mu_{ir}) = \mu_{ir}(1-\mu_{ir})$ and $Corr(Y_i)$ is the correlation matrix.

Let the survey population of size M with S individuals be selected using a stratified multistage sampling design. Let h be the strata (h=1, 2,…., H), k be the cluster (k=1, 2, …., K_h), i denote the individual's index and ω_{hki} denote the longitudinal weight to the i^{th} individual in the k^{th} cluster from the h^{th} stratum. The survey GEE estimator proposed by Rao is the solution of the following equation:

$$\hat{U}_{GEE}\left(\vec{\beta}\right) = \sum_{hki \in S_l} \omega_{hki} D_{hki}{}^{T}(\vec{\beta}) A_{hki}{}^{-1/2}(\vec{\beta}) \hat{R}^{-1} A_{hki}{}^{-1/2}(\vec{\beta}) \left(Y_{hki} - \mu_{hki}(\vec{\beta})\right) = 0 \qquad (3.20)$$

where S_l denotes the longitudinal sample.

The regression parameter estimators $\hat{\vec{\beta}}_{GEE}$ can be obtained from the survey GEE, which takes into account the effects of complex survey design. The generalized estimating equations (GEE) have no closed form with the non-identity link function (i.e., logit for binary response). Iterative methods are required to determine the regression coefficients $\hat{\vec{\beta}}_{GEE}$. The iterative procedure can be defined as

$$\hat{\vec{\beta}}_{GEE}{}^{(k)} = \hat{\vec{\beta}}_{GEE}{}^{(k-1)} - \left(\frac{\partial \hat{U}_{GEE}}{\partial \hat{\vec{\beta}}} \right)^{-1} \left(\hat{\vec{\beta}}_{GEE}{}^{(k-1)} \right) \hat{U}_{GEE} \left(\hat{\vec{\beta}}_{GEE}{}^{(k-1)} \right) \tag{3.21}$$

Estimated regression coefficients $\hat{\vec{\beta}}_{GEE}$ obtained by the GEE approach are consistent even if the covariance structure is selected incorrectly [1].

3.2.1.1.2 Variance estimation

(i) Sandwich variance estimators

As mentioned above, the standard errors of the regression coefficients $\hat{\vec{\beta}}$ play a major role in determining the p-value and confidence interval (C.I.) of $\hat{\vec{\beta}}$. The p-value and the confidence intervals are used to test the null hypotheses. Therefore, unbiased standard errors of $\hat{\vec{\beta}}$ are necessary to make a valid inference. The sandwich estimator is commonly used to estimate the variance of the regression coefficients $\hat{\vec{\beta}}$. This approach provides valid standard errors (SE) of $\hat{\vec{\beta}}$ even if the models and the covariance structure are specified incorrectly [1, 14, 38]. The sandwich estimator can be expressed as

$$Cov(\hat{\vec{\beta}}) = \sum_{i=1}^{n} \left(D_i{}^T V_i^{-1} D_i \right)^{-1} \left(\sum_{i=1}^{n} D_i{}^T (Y_i - \mu_i)(Y_i - \mu_i)^T D_i \right) \left(\sum_{i=1}^{n} D_i{}^T V_i^{-1} D_i \right)^{-1} \tag{3.22}$$

where $D_i = \frac{\partial \mu_i}{\partial \vec{\beta}}$, $V_i = A_i^{\frac{1}{2}} Corr(\mathrm{Y}_i) A_i^{\frac{1}{2}}$, and $\mu_i = \mathrm{E}(\mathrm{Y}_{ir} = 1 | \mathrm{X}_{ir})$.

3.2.2 Multilevel modeling–scaled weights (MM-SW)

Longitudinal data can be treated as two-level clustered data in which the repeated measurements are nested within individuals. The individuals are clusters, and the repeated measurements are units within the cluster in longitudinal data. The clusters or individuals can be nested within super clusters in longitudinal studies, which can be treated as three-level clustered data. For example, in the National Population Health Survey (NPHS), the repeated measurements (units) of the responses are nested within individuals and individuals are nested within primary sampling units (PSU = super cluster). The repeated measurement of responses within the same PSU may be correlated and may be more correlated within an individual for the same PSU.
The statistical methods chosen for analysis should take into account within-cluster dependence when analyzing longitudinal data. We have already discussed multilevel (two-level) modeling for cross-sectional complex survey data in section 3.1.1.2. The additional feature of longitudinal data is the repeated measurements of responses for the same individual, making it three-level clustered data.

Let us consider the dichotomous (yes, no) response variable y_{rik} for three-level clustered data (longitudinal) that is measured at the lowest level in the hierarchical data structure and let x_{rik} be the explanatory variable for the rth unit in level 1 within the ith unit in level 2 within the kth unit in level 3. Three-level generalized linear mixed models with linear predictors can be defined as follows [25]:

$$\mathrm{Logit}\left[pr\left(y_{rik} = 1 | \vec{x}_{rik}, \zeta_{ik}^{(2)}, \zeta_k^{(3)}\right)\right] = \vec{x}_{rik}\vec{\beta} + \vec{x}_{rik}\zeta_{ik}^{(2)} + \vec{x}_{rik}\zeta_k^{(3)}) \qquad (3.23)$$

Let $\zeta_{ik}^{(2)}$ and $\zeta_k^{(3)}$ be the random effects varying over clusters i within super-clusters k and over super-clusters k, respectively.

The following two types of multilevel models are commonly used:

i) Multilevel logistic random-intercept models.

ii) Multilevel logistic random-coefficients models.

3.2.2.1 Multilevel logistic random-intercept models

The multilevel (three-level) logistic random-intercept models, in which units r (level 1) are nested within clusters i (level 2) and clusters i (level 2) are nested within super-clusters k (level 3), can be defined as follows:

$$\text{Logit}\left[\Pr\left(y_{rik}=1 \mid \vec{x}_{rik}, \zeta_{ik}^{(2)}, \zeta_{k}^{(3)}\right)\right]$$

$$= \nu_{rik} = \vec{x}_{rik}\vec{\beta} + \zeta_{ik}^{(2)} + \zeta_{k}^{(3)}$$

$$= \left(\beta_0 + \zeta_{ik}^{(2)} + \zeta_{k}^{(3)}\right) + \beta_1 x_{1rik} + \beta_{2rik} x_{2rik} + \ldots + \beta_p x_{prik} \qquad (3.24)$$

where $\zeta_{ik}^{(2)} \mid x_{rik}, \zeta_{k}^{(3)}, \sim N(0, \psi^{(2)})$ are varying over level 2 within level 3 and $\zeta_{k}^{(3)} \mid x_{rik} \sim N(0, \psi^{(3)})$ are varying over level 3 . The model assumes that the random effects $\zeta_{ik}^{(2)}$ and $\zeta_{k}^{(3)}$ are independent of each other and across clusters, and $\zeta_{ik}^{(2)}$ is also independent across units. For example, in the National Population Health Survey (NPHS) longitudinal data, repeated measurements r (level 1 units) are nested within-individual and individuals i (level 2 units) are nested within-PSU k (level 3 units) .

3.2.2.2 Multilevel modeling random-coefficient models

With the same notation, the multilevel (three-level) logistic random-coefficient model can be defined as

$$\text{Logit}\left[\Pr\left(y_{rik}=1 \mid \vec{x}_{rik}, \zeta_{ik}^{(2)}, \zeta_{k}^{(3)}\right)\right]$$

$$= \nu_{rik} = \vec{x}_{rik}\vec{\beta} + \vec{x}_{rik}^{(2)}\zeta_{ik}^{(2)} + \vec{x}_{rik}^{(3)}\zeta_{k}^{(3)}$$

$$= \left(\beta_0 + \zeta_{ik}^{(2)} + \zeta_k^{(3)}\right) + \left(\beta_1 + \zeta_{ik}^{(2)} + \zeta_k^{(3)}\right) x_{1rik} + \left(\beta_2 + \zeta_{ik}^{(2)} + \zeta_k^{(3)}\right) x_{2rik} + \ldots + \left(\beta_p + \zeta_{ik}^{(2)} + \zeta_k^{(3)}\right) x_{prik}$$

(3.25)

3.2.2.3 Multilevel pseudo maximum likelihood (MPML)

The multilevel pseudo maximum likelihood (MPML) function for three-level clustered data is an extension of the multilevel pseudo maximum likelihood function for two-level clustered data. Let us consider the three-stage sampling design for longitudinal complex survey data, where π_k (k=1, 2,…, $n^{(3)}$) is the probability of selecting a level 3 unit, $\pi_{i|k}$ (i= 1, 2, …, $n^{(2)}$) is the probability of selecting a level 2 unit within level 3, and $\pi_{r|i,k}$ (r=1, 2, ….., $n^{(1)}$) is the probability of selecting the r^{th} unit in level 1 within the i^{th} cluster in level 2 within the k^{th} super-cluster in level 3. Let $\omega_k = 1/\pi_k$, $\omega_{i|k} = 1/\pi_{i|k}$ and $\omega_{r|i,k} = 1/\pi_{r|i,k}$ be the inverse probability of selecting the k^{th} unit in level 3, the i^{th} unit in level 2 within the k^{th} unit in level 3, and the r^{th} unit in level 1 within the i^{th} unit in level 2 within the k^{th} unit in level 3, respectively. The multilevel pseudo maximum log-likelihood function can be defined [6] as

$$L(\vec{\beta}) = \sum_{k=1}^{n^{(3)}} \omega_k \log \int_{-\infty}^{+\infty} \exp\left[\sum_{i=1}^{n^{(2)}} \omega_{i|k} \log \int_{-\infty}^{+\infty} \exp\left\{\sum_{r=1}^{n^{(1)}} \omega_{r|ik} \log f\left(y_{rik} \mid \zeta_{ik}^{(2)}, \zeta_k^{(3)}\right)\right\} g^*\left(\zeta_{ik}^{(2)}\right) d\zeta_{ik}^{(2)}\right] g\left(\zeta_k^{(3)}\right) d\zeta_k^{(3)},$$

(3.26)

where $g^*(\zeta_{ik})$ and $g(\zeta_k)$ are the normal density functions of ζ_{ik} and ζ_k, respectively, and $\vec{\beta}$ is the vector of the parameters.

3.2.2.4 Scaling of weight

The scaling of weight methods were discussed in Section 3.1.2.3.4 in detail. The same scaling methods will be used to scale the weight for longitudinal complex survey data.

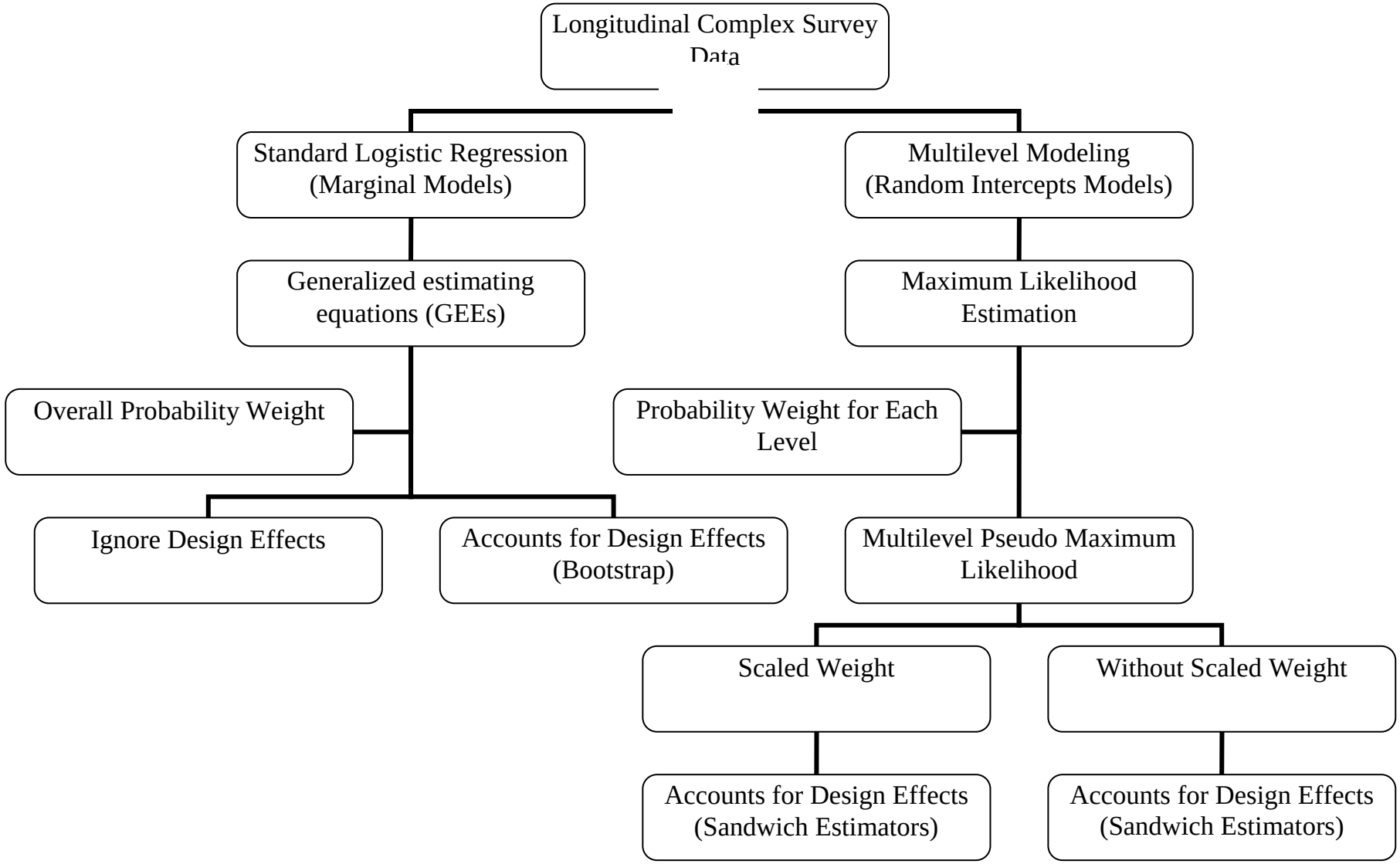
3.2.2.5 Variance estimation

The sandwich estimator technique is used to estimate the standard errors (SE) in multilevel modeling when analyzing longitudinal complex survey data. We defined the sandwich estimator technique in section 3.1.2.4.1 in detail.

3.2.2.6 Goodness-of-fit test

Akaike's Information Criterion (AIC) is commonly used as a model-selection criterion. AIC is based on the maximum likelihood estimate (MLE) in which the distribution of the outcome is known. Generalized estimation equations (GEE) are based on the quasi-likelihood method. Therefore, AIC is not appropriate to use as a model-selection criterion for GEE models utilized to analyze longitudinal complex survey data. Wei Pan proposed the QIC (quasi-likelihood under independent criterion) method as a goodness-of-fit test for models based on the GEE approach [82]. The QIC method can be used for selection of any general working correlation structure based on quasi-likelihood. The QIC(I) are constructed based on quasi-likelihood under the working independent correlation structure (I), and the $QIC_u(R)$ [$QIC_u(R) \equiv -2Q(\hat{\beta}(R); I, D) + 2p$, where p is the number of parameters in the model] are constructed based on quasi-likelihood under a general working correlation structure (R) other than the independent structure. For model selection, the smaller values of $QIC_u(R)$ indicate that the models fit the data adequately. If the $QIC_u(R)$ approximates the QIC, then the GEE model fits the observed data perfectly.

Figure 3.2 Flow chart of statistical methods to accomplish Objective 2



3.3 Statistical methods to accomplish Objective 3

Statistical methods are developed based on certain theoretical assumptions. The efficiency and the power of statistical methods depend on those theoretical assumptions. If the data meet all the theoretical assumptions, then statistical methods can provide a valid and efficient estimation of the parameters estimates from analyses of the survey data [103]. If the data do not meet the assumptions, then the validity of the parameters estimates is not guaranteed. Consequently, the inferences will be invalid based on the analyses of such data. Complex survey data, including cross-sectional and longitudinal data, have many complicated features, such as stratification, clustering and unequal probability selection of sampling units. It is necessary to take into account the effects of these design features to estimate the reliable parameter estimates from the complex survey data. The MM-SW technique and the SR-RV estimation technique are frequently used to analyze complex survey data and, these both statistical techniques are taking into account the complicated features of complex survey data.

The aim of the third objective of my thesis is to investigate which statistical method is appropriate to analyze the cross-sectional complex survey data. The Monte Carlo simulation study is frequently used to assess the power of statistical methods which might be the best tool for comparing the performance of these two statistical methods. It is not possible to assess the performance of statistical methods from the analysis of single real-life data due to some limitations. As a result, computer-based simulated data might be the best choice for assessing the performance of statistical methods. The generation of random numbers is the main part of a simulation study. The availability of statistical software has increased the utilization of Monte Carlo simulation studies. My primary goal is to compare the following two statistical techniques:

(1) Multilevel modeling–scaled weights technique and

(2) Standard regression–robust variance estimation technique

based on analysis of the simulated cross-sectional complex survey data. These two statistical techniques are also used to analyze the real life Canadian Heart Health Survey (CHHS) data to assess and compare the performance of these two statistical techniques. A Monte Carlo simulation study was conducted to generate simulated data with 100 and 1000 number of replications using SAS® software program and the sample size of each data set is 1,731. Both statistical techniques were applied to analyze each of these simulated data, and the two statistical techniques were compared based on the results of the analyses of these simulated data.

3.3.1 Monte Carlo Simulation Technique

Generating random numbers and simulating samples of random variables from a given probability distribution are the main parts of the simulation study [62]. Random numbers are generated by SEED in the SAS® software program. SEED can be defined as follows: a non-negative pseudo-random integer with values less than $2^{31}-1$, generated by the random number function and call routines, is called SEED. SEED is necessary to execute the call routine [65, 67].
After generating the sequences of random numbers, these random numbers are transformed to simulate a sample of random variables with the given probability distribution. The RANTBL functions are used in the SAS ® software program to simulate a sample of random variables from the given probability distribution for a categorical variable. The RANTBL functions are defined as follows.

The RANTBL (SEED, P_1, P_2, P_3, …, P_n, X) function updates SEED and generates a random variable from the probability mass function using the given probability P_1, P_2,…., P_n. The inverse transformation method is used to simulate the discrete probability distribution of a probability mass function. The probability mass function for i^{th} random samples can be defined as

$$f(i)=\begin{cases} P_i & \text{for i=1, 2, 3,……n} \\ 1-\sum_{i=1}^{n} P_i & \text{for i = n+1} \end{cases} \qquad (3.27)$$

where $\sum_{i=1}^{n} P_i \le 1$, $P=(P_1,P_2,...,P_n)$ is a vector of probabilities, and n is the largest integer such that n is less than or equal to the size or dimension of P.

3.3.2 Monte Carlo Simulation Technique using the CHHS

The Canadian Heart Health Survey (CHHS) is a cross-sectional complex survey data which was conducted on 1986-1992 among ten Canadian provinces. The total numbers of participants were 23,129 from ten Canadian provinces. For simplicity, the data only for the Saskatchewan province from the Canadian Heart Health Survey were used to conduct the Monte Carlo simulation study for this thesis. The sampling design for the simulated data was similar to the Canadian Heart Health Survey for the Saskatchewan province. The data collection procedures and sampling design for CHHS were discussed in detail in Section 4.1.

The sample size for the data obtained from the Saskatchewan province was 1,731. Hence, the sample size for each of the simulated datasets was 1,731. The response variable of interest was the type 2 diabetic status (yes, no), where "yes" means people who had type 2 diabetes and "no" means people who did not have type 2 diabetes. The covariates were body mass index (<25 kg/m^2, ≥25 kg/m^2) and education level (<secondary, >= secondary) for this simulation study.
The primary sampling unit (PSU), the probability weight and 500 bootstrap weights for each participant were available in CHHS. The stratification based on area of residence (rural, urban), sex (male, female), age group (18–44years, 45–64 years, and 65 years and above) and PSU level were used to calculate probability weight and 500 bootstrap weights in the CHHS. These probability

weight and 500 bootstrap weights were computed by methodologists in statistics Canada. The combination of area, sex, age group, and PSU level were used to generate the simulated data according to Saskatchewan data. The probability weight for each participant from the CHHS was used with the combination of area, sex, age group and PSU level for each of the simulated data sets. The 500 bootstrap weights for each participant in the Saskatchewan data were also used with the combination of area, sex, age group and PSU level for each of the simulated data sets. The detail generating procedures of simulated data based on Saskatchewan data were discussed in section 4.3. The outcome variable of interest was type 2 diabetic status (yes, no) , and the covariates were body mass index (<25 kg/m^2, ≥ 25 kg/m^2) and education level (<secondary, $\geq$ secondary). The covariates were generated randomly using the SEED and RANTBL functions in the SAS® software program. In order to generate the simulated data using the RANTBL function in the SAS® software program, the proportion of each category of each covariate was required. The proportion for each category of each variable was obtained from the Saskatchewan data to generate the simulated data. The SEED and RANTBL functions were discussed in section 3.3.1.

A total of 100 and 1000 Monte Carlo simulated data sets were generated based on the above setup in SAS® software program separately. The reasons of generating two groups (100 and 1000) of simulated data with different numbers of replications were to compare the performance of MM-SW technique and SR-RV estimation technique as well as to determine the effects of number of simulations on parameter estimates. The number of replications is one of the key criteria of the Monte Carlo simulation study. It is commonly known that the higher number of replications provide usually consistent and precise parameters estimates [100]. To my knowledge, a limited number of studies have provided the formula or general criteria to determine the number of simulations/replications required for a simulation study. Burton et al. demonstrated the following formula to calculate the approximate number of replications for a simulation study [99].

$$B = \left(\frac{Z_{1-\alpha/2}\sigma}{\delta} \right)^2 \quad (3.28)$$

where δ denotes the level of accuracy, σ^2 denotes the variance of regression parameter and $Z_{1-\alpha/2}$ is the $1-\alpha/2$ quintile of the standard normal distribution. This formula was used to determine the approximate number of replications for this simulation study. The standard regression–robust variance estimation technique and the multilevel modeling–scaled weights technique were applied to analyze each simulated dataset in STATA. The scaled weight was used in multilevel modeling which was discussed in section 3.1.2.3.4. Both statistical techniques were used to analyze two groups of simulated datasets separately and evaluated based on following criteria (Table 3.3.1).

Table 3.3.1: The statistical formula for the assessment criteria [99]

Assessment criteria	Formula
Bias of regression coefficient	$\delta = \overline{\hat{\beta}}_{simu} - \beta_{true}$
Relative or percentage bias of regression coefficient	$\left(\frac{\overline{\hat{\beta}}_{simu} - \beta_{true}}{\beta_{true}} \right) * 100$
Standardized bias of regression coefficient	$\left(\frac{\overline{\hat{\beta}}_{simu} - \beta_{true}}{SE(\beta_{true})} \right) * 100$
Means square error (MSE) of regression coefficient	$\left(\overline{\hat{\beta}}_{simu} - \beta_{true} \right)^2 + \left(SE(\beta_{simu}) \right)^2$
Coverage of true regression coefficients	Proportion of times the 95% Confidence interval $\hat{\beta}^i_{simu} \pm Z_{1-\alpha/2} SE\left(\hat{\beta}^i_{simu} \right)$ For i=1,2,3,…,B
Average 95% confidence interval length	$\frac{\sum_{i=1}^{B} 2Z_{1-\alpha/2} SE\left(\hat{\beta}_i \right)}{B}$
Relative efficiency	$\text{RE } (\hat{\beta}^{ml}, \hat{\beta}^{sr}) = \frac{\text{var}(\hat{\beta}^{ml})}{\text{var}(\hat{\beta}^{sr})}$

Note: β_{true} and $\overline{\hat{\beta}}_{simu} = \frac{\sum_{i=1}^{B} \hat{\beta}^i_{simu}}{B}$ denote the true value and simulated average value of regression coefficients respectively, where B denotes the number of simulations.

The statistical method with less bias, smaller MSE, narrower length of 95% C.I., higher coverage of the observed or true regression coefficients in the corresponding simulated 95% confidence intervals obtained from the analysis of simulated data can be considered as a better method. Estimation was performed in the standard regression–robust variance estimation technique using the "logit" function with probability weight and 500 bootstrap weights in STATA. Estimation was performed in the multilevel modeling–scaled weights technique by "GLLAMM" with 12-point adaptive quadrature in STATA.

CHAPTER 4

DESCRIPTION OF THE POPULATIONS OF STUDY

4 Introductions

Datasets, especially in public health, are valuable sources of information, and these datasets provide many types of information, such as disease information, area level measurements, disease status, and reasons for the development of disease, which can be very useful for applied researchers. Both the Canadian Heart Health Survey (CHHS) and the National Population Health Survey (NPHS) are huge datasets and provide unique health information about Canadians.
Datasets from the Canadian Heart Health Survey (CHHS) and the National Population Health Survey (NPHS) were used to accomplish Objective 1 and Objective 2, respectively. Detailed descriptions of the CHHS and the NPHS are provided in sections 4.1 and 4.2. The CHHS data set was also used to generate simulated data to accomplish the objective 3 which are revealed in section 4.3.

4.1 Cross-sectional complex survey data: CHHS

The Canadian Heart Health Survey (CHHS) is a population-based survey that was conducted to determine the status of cardiovascular disease (CVD) at the provincial and national levels in Canada [83]. This study was collaboratively conducted by the Heart and Stroke Foundation of Canada, Health Canada and the Provincial Department of Health in each province. The primary objective of this survey was to determine the prevalence of CVD risk factors, the knowledge and awareness levels of CVD causes and the consequences of CVD among Canadian. The CHHS consists of two sets of integrated data: core information collected by all ten provincial surveys and family history (i.e., father, mother, brothers, sisters, etc.) related to heart disease collected by only four provinces—Quebec, Ontario, Saskatchewan and Alberta.

4.1.1 Study design

In the CHHS, a multistage stratified probability sampling design was used to select independent samples at each province in Canada. The survey was conducted among all Canadian provinces, and each province was divided into rural, urban and metropolitan areas. Urban areas were stratified into numbers of urban strata based on their population sizes, and rural areas were stratified by standard geographic areas (e.g., census division, health units), which were called rural stratum. The number of stratum from urban strata and rural strata were selected using probability proportional to size (PPS) in each province. Each of these selected areas was further stratified into six age/sex (male, female, 18–34 years, 35–64 years, 65–74 years) stratum in each province, and independent simple random samples (SRS) were drawn from each stratum. Municipalities, counties, census lots, census districts or health units were defined as primary sampling units (PSU). PSU was selected using probability proportional to size (PPS).

4.1.2 Probability weight

Two probability weights were calculated by statistical methodologist for each participant within each province to adjust for the unequal probability of selection and non-responses at the home interview (PWGTQ) and clinic visit (PWGTC). The PWGTQ probability weights were used for information collected at home interviews and the PWGTC probability weights were used for information collected during clinic visits. The PWGTC probability weights were used for analyses of the information collected jointly during both home interviews and clinic visits. The following formulas were used to calculate these probability weights.

The formula for the probability weight (PWGTQ) for respondent from at home interviews was [83]:

$$\omega_{phair} = \left(P_{pi} \Big/ \hat{P}_{pi} \right) \omega^{*}{}_{phair} \tag{4.1}$$

The formulas for the probability weight (PWGTC) for respondent from both at home interviews and clinic visits was [83]:

$$t_{phair} = \left(P_{pi} \Big/ \hat{P}_{pi} \right) t^{*}{}_{phair} \tag{4.2}$$

where

p - Province, h- stratum, a- area, i – age/sex group, r – number of replicates from age/sex group,

N_{phair} - Number of persons on the medical insurance registers (MIR) of province "p" in stratum "h" area "a" and age/sex group "i";

n_{phair} - Number of persons selected from province "p";

m_{phair} - Number of persons out of (n_{phair}) responded to the home interview;

s_{phair} - Number of persons out of (m_{phair}) came to the clinic;

α_{pha} - First stage selection probability factor for area "a" selected from stratum "h" and province "p";

P_{pi} - Statistics Canada population estimates (closest to the survey date) of province "p" by age/sex "i";

$\hat{P}_{pi}$ - Estimate of P_{pi} from the survey;

r - Number of replicates selected from age/sex group "i";

$\omega^{*}{}_{phair} = \alpha_{pha} N_{phai} / \left(m_{phair} \right)$

$$t^{*}_{phair} = \alpha_{pha} N_{phai} / \left(s_{phair} \right)$$

$$\hat{P}_{pi} = \sum_{h,a} \alpha_{pha} N_{phai}$$

4.1.3 Study population

All male and female participants (23,129), aged 18–74 years, were recruited from ten Canadian provinces. In this thesis, the total 21,021 participants from nine Canadian provinces (except Nova Scotia) were included. The reason for excluding Nova Scotia was that the location of the residence for participants from this province was not recorded. The people who were living on Indian reserves, in military camps, and in institutions such as prisons were excluded from this survey. Participants who moved to a new address within the same area were included in the survey.

4.1.4 CHHS data collection

The CHHS data were collected into two phases, using the medical insurance registers (MIR) as a sampling frame from each province. In the first phase, participants were visited into their homes by public health nurses and interviewed with a questionnaire. This questionnaire collected information on cardiovascular disease (CVD) risk factors, on attitudes and opinions about heart health risk factors including basic demographic characteristics and on lifestyle (smoking, physical activity, alcohol intake). Information was also collected on chronic disease status, such as diabetic status and hypertensive status. Two blood pressure readings were taken at the time of the interview, one at the beginning and the other at the end of the interview. In the second phase, participants who were interviewed at home were invited to visit a clinic within two weeks after the home interview. After at least eight hours of fasting, blood pressure (systolic and diastolic), blood samples, and anthropometric measurements were obtained at the clinic visit. The total number of respondents

who were attended both the home and clinic visits was 23,129 from ten Canadian provinces. The response rates were approximately 77% for the home visit and 67% for both the home and clinic visits [83].

Table 4.1 Sample size stratified by province

Province	Number of persons located	Number of persons interviewed at home	Number of persons who visited a clinic
Newfoundland	3185	2394	2067
Prince Edward Island	2318	2088	2026
Nova Scotia	2735	2108	1798
New Brunswick	2737	2093	1948
Quebec	3052	2353	2095
Ontario	3639	2538	2039
Manitoba	3597	2766	2316
Saskatchewan	2893	2158	1749
Alberta	2739	2237	1993
British Columbia	2960	2394	2064
Total	29,855	23,129	20,095

4.1.5 Outcome variable of interest

The outcome variable of interest for our study was self-reported type 2 diabetes diagnosed by a physician or health-care professional. The outcome variable was a dichotomous (yes, no)

variable, where “yes” indicated a positive response and “no” indicated a negative response to the following question: “Have you ever been told by a doctor that you have diabetes?”

4.1.6 Risk factors of type 2 diabetes

Based on the literature review, the potential risk factors that were available in the CHHS dataset and were considered for analysis were age in years, sex, location of residence, level of education, household income per year, marital status, employment status, physical activity, and body mass index. Detailed definitions of these variables are included below.

Age in years:

This variable indicates the age in years during the home interview of each participant in this study. The age variable was a continuous variable, and it was divided into three categories (18–44 years, 45–64 years, and 65–74 years). Type 2 diabetes usually develops after the age of 40 and increases among older people. These age categories were made based on the literature review.

Sex: The sex of the participants was known from the demographic information.

Location of residence (rural/urban):

The location of the residence was a derived variable that was determined using the definitions of rural and urban areas provided by the Statistics Canada. Areas where 1000 or fewer people lived were called rural, and areas where more than 1000 people lived or the population density was 400 or more per square kilometer were called urban.

Level of education:

The education level variable represents the level of education for each participant in the CHHS. It was a categorical, derived variable and was further recoded into three categories for our

analysis: elementary (no schooling, elementary and some secondary), secondary (secondary school graduation, other post-secondary, some community college, and diploma/certificate: trade school), university (some university, diploma/certificate/CEGEP, bachelor's degree, master's/medicine/doctorate). The following question was asked: "What is the highest grade or year of school you have completed?"

Household income per year:

The household income variable described the total household income for each participant. It was a categorical variable and the categories were as follows: <$12,000; $12,000–$24,499; $25,000–$49,999; and >$50,000 per year.

Employment:

The employment variable illustrated the employment status of each participant. This was a categorical variable with the following categories: full time (35 hours or more a week), part time (less than 35 hours a week)/student, unemployed/laid off, homemaker, and retired.

Physical activity:

The physical activity variable described whether or not the participants were involved in any physical activity once or more per week. The questionnaire for all provinces except Saskatchewan was similar. The physical activity variable was a categorical variable (yes, no), where "yes" meant the person engaged in physical exercise at least once a week and "no" meant the person did not engage in physical exercise. The following question was asked: "Do you regularly engage in physical exercise during your leisure time? By regularly, we mean at least once a week during the past month".

Body mass index:

The body mass index (BMI) variable was determined from the height and weight of each participant. It was calculated based on the following formula:

$BMI = \frac{\text{Weight in kilograms}}{\text{Height in meters}^2}$. The BMI was a continuous variable and was categorized into three groups for our analysis: normal weight (BMI<25 kg/m^2), overweight (BMI≥ 25.0 kg/m^2 and < 30.0 kg/m^2) and obese (BMI≥30 kg/m^2) [84].

4.2. Longitudinal complex survey data: National Population Health Survey (NPHS)

The National Health Information Council (HNIC) first proposed conducting an ongoing national population health survey among Canadian populations in 1991, and Statistics Canada received funding to conduct this survey based on this recommendation in 1992.

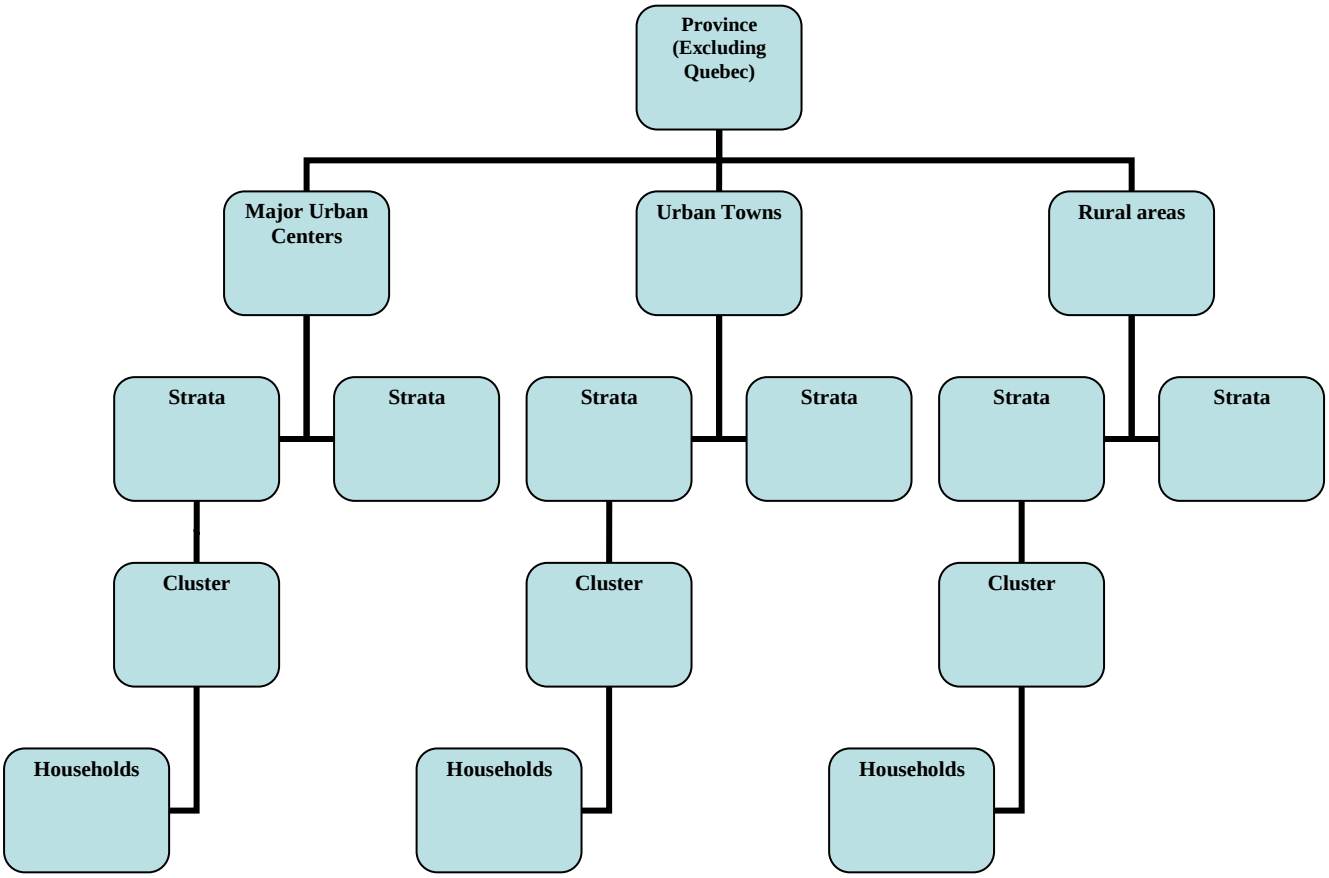
The NPHS was a cohort study consisting of longitudinal complex surveys on the same population that was commenced in 1994/95 by Statistics Canada [85]. It will continue every two years until 2014. Longitudinal information on the health of the Canadian population and socio-demographic information was collected on those people who were selected in Cycle 1 (1994/95). The survey design for the NPHS was formed based on the Labor Force Survey (LFS) design. The questionnaire addressed health status, use of health services, determinants of health, chronic conditions, activity restrictions, and socio-demographics such as age, sex, education, household income, and labor force status.

4.2.1 Study design

A stratified multistage sampling design was used to conduct the national population health survey (NPHS). The same sampling design was used for each province except Quebec. In the first stage, each province was divided into three areas—major urban centers, urban towns and rural areas—and homogeneous strata were formed from each separate geographic and/or socio-economic stratum. The independent samples of clusters were selected using probability proportional to size

(PPS) from each stratum. Six clusters were selected from each stratum. In the second stage, households were selected from the list of dwellings that was prepared from each selected cluster. In Quebec, the NPHS sample was selected from dwellings participating in a Santé Québec health survey in 1992/93, Enquête social et de santé (ESS). The survey sampled 16,010 dwellings using a two-stage sample design similar to other provinces. The province was divided geographically into 15 health areas with four urban classes: Montreal Census Metropolitan Area, regional capitals, small urban agglomerations and the rural sector. In each area, clusters were stratified by socio-economic characteristics and selected using PPS sampling. Samples of dwellings were randomly drawn from each cluster.

Figure 4.1 The survey methodology for NPHS data.



4.2.2 Study population

The target population of the household component included all residents in the ten Canadian provinces in 1994–95. People who were living on Indian Reserves and Crown lands, who were residents of health institutions, were full-time members of Canadian Forces Bases and were in remote areas of Ontario and Quebec were excluded.

The sample size of the longitudinal NPHS data was 17,276, with participants aged 12 to 99 years. No new participant was included after 1994–95. The sample size of the longitudinal study by province in 1994–95 and the number of participants that provided a full response to all six cycles are shown in Table 4.2.

Table 4.2 Longitudinal sample size stratified by province

Province	Longitudinal Sample Cycle 1(1994–95)	Number of respondents providing a full response in Cycles 1, 2, 3, 4, 5 and 6
Newfoundland	1,082	768
Prince Edward Island	1,037	746
Nova Scotia	1,085	732
New Brunswick	1,125	758
Quebec	3,000	1,969
Ontario	4,307	2,733
Manitoba	1,205	868
Saskatchewan	1,168	870
Alberta	1,544	1,033
British Columbia	1,723	1,116
Total	17,276	11,593

Note: Cycle1 = 1994-95, Cycle2= 1996-97, Cycle3 = 1998-99, Cycle4 = 2000-01,

Cycle 5 = 2002-03, Cycle 6 = 2004-05,

4.2.3 NPHS data collection

The NPHS collected socio-demographic information, such as age, sex, education, household income, labor force status, health status and use of health services. The NPHS started to conduct this survey initially with 19,600 households, with a minimum of 1,200 households in each province. The longitudinal survey did not include people who immigrated to Canada after 1994–1995. The NPHS has completed six cycles to date that are available for public use. The cycles are Cycle 1 (1994–95), Cycle 2 (1996–97), Cycle 3 (1998–99), Cycle 4 (2000–01), Cycle 5 (2002–03) and Cycle 6 (2004–05).

All participants aged 18 years and older were included in this study. This study was conducted based on questionnaires that were designed for computer-assisted interviewing (CAI). Participants were contacted by telephone. Proxy reporting was allowed for respondents who were less than 12 years of age; proxy reporting for those over 12 was allowed only for reasons of illness or incapacity. The response rates of 17,276 panel members for each cycle are shown in Table 4.3.

Table 4.3 Response rate for each cycle

Cycle	Response Rate
Cycle1	86.0%
Cycle2	93.6%
Cycle3	88.9%
Cycle4	84.8%
Cycle5	80.6%
Cycle6	77.4%

4.2.4 Probability weight

Unequal probabilities of selection and non-response are the common features of longitudinal complex survey data. Weighted data must be used to obtain valid estimates of parameters in complex survey data. The main role of weighting in complex survey data such as the NPHS is that each individual in the sample represents other individuals, including him- or herself. The estimate of parameters based on complex survey data cannot be meaningful without weighting. Weighting in longitudinal survey data represents the inverse of probability of selection of the individual analysis at the time of sample selection. In the NPHS, weighting represents the inverse of probability of selection of an individual who took part in cycle 1 (1994–95) but not in subsequent cycles. The probability weight in the NPHS was obtained by the post-stratifying cycle 1 stripped weights for the 1994–95 population estimates based on a 1996 census count by age groups (0–11, 12–24, 25–44, 45–64, 65 and older) and sex within each province. The post-stratification adjustment is given by the following ratio [85]:

$$\frac{\text{Population estimate in a province/age/sex category}}{\text{Sum of “stripped” weights of respondent household numbers in a province/age/sex category}}$$

4.2.5 Outcome variable of interest

The outcome variable of interest in our study was self-reported, professionally diagnosed type 2 diabetes. The outcome variable was a dichotomous variable (yes or no). The following question was asked of the participant: “Do you have any of the following long-term conditions that have been diagnosed by a health professional? – Diabetes”. Here, “yes” indicated a positive response to this question, and “no” indicated a negative response to this question.

4.2.6 Risk factors for type 2 diabetes

The possible covariates of our study, including confounding, were the following: age, sex, area, body mass index (BMI), education level, household income, physical activity, family history of type 2 diabetes (father or mother has diabetes), and cycle. These were expected to be independent risk factors for type 2 diabetes. All of these variables were available in the NPHS datasets. Confounders and effect modifiers for type 2 diabetes were also examined during the analysis.

Age in years:

This is a continuous variable that was collected every cycle during the interview period. The study population in our analysis included panel members who were 18 years and older at each cycle. The age variable was categorized into the following groups: 18–44 years, 45–64 years, and 65–75 years.

Sex:

The sex (male, female) of all participants in the NPHS data was known.

Area (rural/urban)—Place of residence:

Rural areas were defined as the areas where few than 1,000 people lived. Urban areas were defined as the areas where more than 1,000 people lived and the population density was 400 or more per square kilometer. The urban areas included the urban core, the urban fringe and the urban area outside the census metropolitan area (CMA). The rural areas included the participants staying in a rural fringe or a rural area outside the CMAs.

Body mass index (BMI):

The body mass index was calculated based on the following formula:

$BMI = \frac{\text{Weight in kilograms}}{\text{Height in meters}^2}$, where height and weight were self-reported. Participants with a height of three feet or less or more than seven feet were excluded from this BMI calculation. The baseline BMI was a continuous variable and was categorized into three groups for our analysis: normal weight (BMI<25 kg/m^2), overweight (25.0 kg/m^2 ≤BMI < 30.0 kg/m^2) and obese (BMI≥30 kg/m^2). These categories of BMI were made based on the Canadian guidelines for body weight classification in adults [86].

Household income per year:

The household income variable represents the total household income from different sources of earning per year. It was a categorical based on derived variable. This variable was recoded into four categories for our analysis: lowest income (0–$14,999/year), lower middle income ($15,000–$29,999/year), middle income ($30,000–$49,999/year) and high income (≥$50,000/year).

Education:

The education variable represents the level of education for each participant. It was a categorical, derived variable and was further recoded into four categories for our analysis: elementary (no schooling, elementary and some secondary), secondary (secondary school graduation, other post-secondary, some community college, and diploma/certificate: trade school), bachelor's degree and higher (some university, diploma/certificate/CEGEP, bachelor's degree, master's/medicine/doctorate). The following question was asked: "What is the highest level of education that you have attained?"

Marital status:

This was a categorical variable indicating the present marital status for each participant, and it was further recoded into three categories: widowed/separated/ divorced, never married/single,

and married/common law/living together. The following question was asked: “What is your current marital status?”

Physical exercise:

This variable represents the frequency of all physical activities lasting more than 15 minutes. The categories of the variable were infrequent, occasion and regular.

Father had type 2 diabetes:

This variable represents the diabetic history of the father of the participants. The following question was asked of the participants: “Did your birth father ever have diabetes?” Father had diabetes (yes or no), where “yes” indicated a positive response and “no” indicated a negative response to this question.

Mother had type 2 diabetes:

This variable represents the diabetic history of the mother of the participants. The following question was asked of the participants: “Did your birth mother ever have diabetes?” Mother had diabetes (yes, no), where “yes” indicated a positive response and “no” indicated a negative response.

4.3 Simulated data for Monte Carlo simulation technique

For the Monte Carlo simulation study, it was required to generate the cross-sectional complex survey data to accomplish the third objective that was to assess the performance of the multilevel modeling-scaled weights technique and the standard regression-robust variance estimation technique to analyze the cross-sectional complex survey data based on Monte Carlo simulation study. The sampling design for the simulated cross-sectional complex survey data was similar to that of the sampling design of Saskatchewan survey. The complex survey design factors such as stratification, clustering, and unequal probability of selection have a significant effect on

parameter estimates. The CHHS is a cross-sectional complex survey which was conducted among ten Canadian provinces. For simplicity I used only Saskatchewan data (a part of the CHHS) with sample size 1,731 for the simulation study.

For this simulation study, the outcome variable was type 2 diabetes (yes, no) where 'yes' means participants who had type 2 diabetes and 'no' means participants who did not have type 2 diabetes and only two independent variables body mass index(<25 kg/m^2, ≥25 kg/m^2) and education level (<secondary, ≥secondary) were used.

It was necessary to determine the approximate number of replications for generating the simulated data. The Monte Carlo simulation technique was based on a real life complex survey Saskatchewan data. The estimated regression coefficients and their standard errors obtained from the analysis of the observed Saskatchewan data were used to determine the number of replications using the equation (3,27). The estimated regression coefficients and their standard errors for two covariates (body mass index (<25 kg/m^2, ≥25 kg/m^2) and education level (<secondary, ≥secondary) were shown in the following table 4.4. The calculated 5% accuracy (δ) of corresponding regression coefficients (e.g. 5%*0.75945=0.0379725 for BMI) are also shown in Table 4.4.

Table 4.4 The estimated number of replications based on observed parameter estimates and their standard errors and expected accuracy.

Variables	Parameter estimates (SE) $(\beta_{true}(\sigma))$	5% accuracy (δ)	Number of simulations (B)
Body mass index (BMI)			
<25 kg/m^2 (ref)	0.75945 (0.2073)	0.0379725	115
≥25 kg/m^2			
Education level			
<secondary	0.4374059(0.3000)	0.0218703	723
≥secondary (ref)			

Based on the calculation of the approximate numbers of replications using the formula given in equation (3.27), 115 replications were required based on the body mass index (<25 kg/m^2, ≥ 25 kg/m^2) variable and 723 replications were required based on the education level (<secondary, ≥secondary) variable to conduct a Monte Carlo simulation. Using these reflections I decided to generate two groups of simulated data sets, one with 100 replications and the other with 1000 replications. The simulated cross-sectional complex survey data sets with the 100 and 1000 numbers of replications were generated using the following steps:

4.3.1 Data set used for Monte Carlo simulation technique

Saskatchewan data were extracted from the complete CHHS data set. The Saskatchewan data was sorted by area (rural, urban), sex, age groups (18-44 yrs, 45-64 yrs, 65 yrs and above) and PSU level (six levels). There were six PSU level in the Saskatchewan data. In order to create the weight file for only weight variables from Saskatchewan data, we sorted this data set by the combination of above variables. The probability weight and 500 bootstrap weights variables were calculated using the combinations of these variables.

4.3.2 Creation of 'weight' data file

A data file called 'weight file' was created only with probability weight variable and 500 bootstrap weights for the Saskatchewan data. A new ID variable was created in the 'weight file' to merge with simulated data sets. The sample sizes for each simulated data sets and Saskatchewan data were same and they had unique identification number for each patient. The 'weight file' with weight variables (probability weight and 500 bootstrap weights) was linked later on with each simulated data set.

4.3.3 Generating simulated data with the combinations of area, sex, age groups and PSU level

In this stage, the following steps were used to generate the simulated data:

1) The proportions of each category of body mass index (<25 kg/m^2, ≥25 kg/m^2) and education level (<secondary, ≥secondary) were estimated with the combinations of area (rural, urban), sex, age groups (18-44 yrs, 45-64 yrs, 65 yrs and above) and PSU levels (six levels) from the Saskatchewan data.

2) There were 44 such combinations of area (rural, urban) sex, age groups (18-44 yrs, 45-64 yrs, 65 yrs and above) and PSU levels (six levels) in the Saskatchewan data

3) Simulated data sets were generated based on each of these combinations using the obtained proportion for each category of each covariate.

4) In order to augment each of these simulated data set with probability weight and 500 bootstrap weights, each of the simulated data sets was linked with 'weight' file by the above combinations. For linkage process (see Figure 4.2).

5) For each of the 44 combinations, RANTBL (SEED, P_1, P_2, …,P_n, X) function in SAS® was used to generate the simulated data (with 100 and 1000 numbers of replications) for body mass index (<25 kg/m^2, ≥25 kg/m^2) and education levels (<secondary, ≥secondary) independently. The logistic regression was used to generate the outcome variable (type 2 diabetes (yes, no)) using body mass index (<25 kg/m^2, ≥25 kg/m^2) and education levels (<secondary, ≥secondary) as independent variables in the model. In this process, first the linear predictor was generated, where initial intercept and initial regression coefficients for body mass index (<25 kg/m^2, ≥25 kg/m^2) and education level (<secondary, ≥secondary) were estimated from the analysis of observed Saskatchewan data using multilevel modeling – scaled weight technique and standard regression-robust variance estimation technique separately, then the inverse link function was used to calculate predicted probability. Finally, the outcome variable of interest (type 2 diabetes) was constructed

using uniform (0,1) distribution. If the random number is less than predicted probability then the observation was defined as '1' otherwise observation was define as '0'. This step was conducted based on two number of replications, One is for 100 replications and other one for 1000 replications.

4.3.4 Creating the simulated data sets with 100 and 1000 number of replications

One hundred simulated datasets, each of size 1,731, were obtained by appending the simulated data sets obtained from each of the 44 combinations with 100 replications. Similarly, One thousand simulated datasets, each of size 1,731, were obtained by appending the simulated data sets obtained from each of the 44 combinations with 1000 replications.

4.3.5 Creating final simulated data sets after linking each simulated data with weight file

After generating the simulated data sets, I merged each of simulated data sets with the 'weight file' which was created for weight variable (see Figure 4.2). The probability weight and 500 bootstrap weights were available in Saskatchewan data. After completion of above steps, the two groups of final simulated cross-sectional complex survey data were created: one with 100 number of simulations and other one with 1000 numbers of simulations.

Figure 4.2 Data linkage between simulated data and weight file created from Saskatchewan data

		Data set 1			Data set 2			… …	Data set 1000		
# of comb.	# of obs.	id	bmi	edu	id	bmi	edu		id	bmi	edu
1	17	1 … … 17			1 … … 17				1 … … 17		
2	13	18 … … 30			18 … … 30				18 … … 30		
3	19	31 … … 49			31 … … 49				31 … … 49		
4	10	50 … … 59			50 … … 59				50 … … 59		
5	64	60 … … 123			60 … … 123				60 … … 123		
6	48	124 … … 171			124 … … 171				124 … … 171		
7	66	172 … … 237			172 … … 237				172 … … 237		

Weight file							
# Of comb.	# of obs.	id	prob. weight	bsw1	bsw2	… …	bsw500
1	17	1 … … 17					
2	13	18 … … 30					
3	19	31 … … 49					
4	10	50 … … 59					
5	64	60 … … 123					
6	48	124 … … 171					
7	66	172 … … 237					

8	58	238 … … 295			238 … … 295				238 … … 295		
9	51	…			…				…		
10	37	…			…				…		
11	47	…			…				…		
12	39	…			…				…		
13	57	…			…				…		
14	36	…			…				…		
15	56	…			…				…		
16	40	619 … … 658			619 … … 658				619 … … 658		
17	67	…			…						
18	45	…			…						
19	73	…			…						
20	45	…			…						
21	111	889 … … 999			889 … … 999				889 … … 999		
22	61	1000 … …. 1061			1000 … …. 1061				1000 … …. 1061		
23	120	1062 … … 1181			1062 … … 1181				1062 … … 1181		
24	56	…			…				…		
25	98	…			…				…		
26	44	…			…				…		

8	58	238 … … 295					
9	51	…					
10	37	…					
11	47	…					
12	39	…					
13	57	…					
14	36	…					
15	56	…					
16	40	619 … … 658					
17	67						
18	45						
19	73						
20	45						
21	111	889 … … 999					
22	61	1000 … …. 1061					
23	120	1062 … … 1181					
24	56	…					
25	98	…					
26	44	…					

27	69	…			…				…		
28	33	…			…				…		
29	19	…			…				…		
30	12	1501 … … 1512			1501 … … 1512				1501 … … 1512		
31	18	1513 … … 1530			1513 … … 1530				1513 … … 1530		
32	10	…			…				…		
33	24	…			…				…		
34	10	…			…				…		
35	20	…			…				…		
36	12	…			…				…		
37	21	…			…				…		
38	10	…			…				…		
39	19	…			…				…		
40	12	…			…				…		
41	19	…			…				…		
42	11	…			…				…		
43	25	1698 … … 1722			1698 … … 1722				1698 … … 1722		
44	9	1723 … … 1731			1723 … … 1731				1723 … … 1731		
		1731			1731				1731		

27	69	…					
28	33	…					
29	19	…					
30	12	1501 … … 1512					
31	18	1513 … … 1530					
32	10	…					
33	24	…					
34	10	…					
35	20	…					
36	12	…					
37	21	…					
38	10	…					
39	19	…					
40	12	…					
41	19	…					
42	11	…					
43	25	1698 … … 1722					
44	9	1723 … … 1731					
		1731					

CHAPTER 5

ANALYSIS OF RESULTS

5.1 MODELS FOR CROSS-SECTIONAL COMPLEX SURVEY DATA

The overall objective of this thesis was to compare the multilevel modeling–scaled weights (MM-SW) technique with the standard regression–robust variance(SR-RV) estimation technique for analyzing cross-sectional and longitudinal complex survey data.

The first objective of this thesis was to compare the MM-SW technique with the SR-RV estimation technique based on an analysis of the cross-sectional Canadian Heart Health Survey (CHHS).

The statistical modeling procedures based on the Canadian Hearth Health Survey (CHHS) using the MM-SW technique and the SR-RV estimation technique are discussed in this chapter.

Characteristics of the study population and their descriptive analyses are described in sections 5.1.1 and 5.1.2, respectively. The estimations of crude prevalence are discussed in section 5.1.3.

The modeling approach for cross-sectional complex survey data (CHHS) and the comparison between the MM-SW technique and the SR-RV estimation technique based on the obtained results are discussed in sections 5.1.4 and 5.1.5, respectively. Interpretations of the empirical results obtained from analyses of the CHHS are discussed in section 5.1.6.

5.1.1 Study Population

The Canadian Heart Health Survey (CHHS) datasets contain 21,021 participants from nine Canadian provinces. All male and female participants, aged 18 to 74 years, from the nine Canadian provinces were included in our analysis. The people living on Indian reserves, in military camps, and in institutions such as prisons were excluded from this survey.

The province of Nova Scotia was not included in the analysis because the variable ‘location of residence (rural or urban)’ was missing for this province. One of the objectives was to compare

the prevalence of self-reported type 2 diabetes (crude and adjusted) between rural and urban residents. The total number of participants, stratified by self-reported type 2 diabetic status, from the nine Canadian provinces is presented in Table 5.1. The response for type 2 diabetic status is based on the question, "Have you ever been told by a doctor that you have diabetes?" The proportions of self-reported, physician-diagnosed type 2 diabetes were highest in Manitoba (7.2%), followed by Alberta (5.5%), Saskatchewan (5.4%), Quebec (5.2%) and Newfoundland (5.1%). Prince Edward Island (3.4%) had the lowest proportion of type 2 diabetes compared with the other provinces.

Table 5.1 Number of participants with type 2 diabetic status in each province

Provinces	Un-weighted		Weighted	
	Type 2 Diabetes		Type 2 Diabetes	
	Yes (%)	No (%)	Yes (%) (95% C.I.)	No (%) (95% C.I.)
Newfoundland	123 (5.1%)	2271 (94.9%)	5.4% (4.8 – 6.2)	94.6% (93.8 – 95.2)
Prince Edward Island	70 (3.4%)	2018 (96.6%)	4.1% (3.1 – 5.5)	95.9% (94.5 – 96.9)
New Brunswick	100 (4.8%)	1993 (95.2%)	5.5% (4.7 – 6.4)	94.5% (93.6 – 95.3)
Quebec	122 (5.2%)	2227 (94.8%)	4.9% (3.9 – 6.2)	95.1% (93.8 – 96.1)
Ontario	112 (4.4%)	2426 (95.6%)	4.0% (3.3 – 4.8)	96.0% (95.2 – 96.7)
Manitoba	200 (7.2%)	2566 (92.8%)	4.9% (4.1 – 5.8)	95.1% (94.2 – 95.9)
Saskatchewan	114 (5.3%)	2044 (94.7%)	5.4% (3.6 – 8.0)	94.6% (92.0 – 96.4)
Alberta	124 (5.5%)	2113 (94.5%)	4.9% (4.3 – 5.6)	95.1% (94.4 – 95.7)
British Columbia	101 (4.2%)	2293 (95.8%)	4.4% (3.9 – 5.0)	95.6% (95.0 – 96.1)

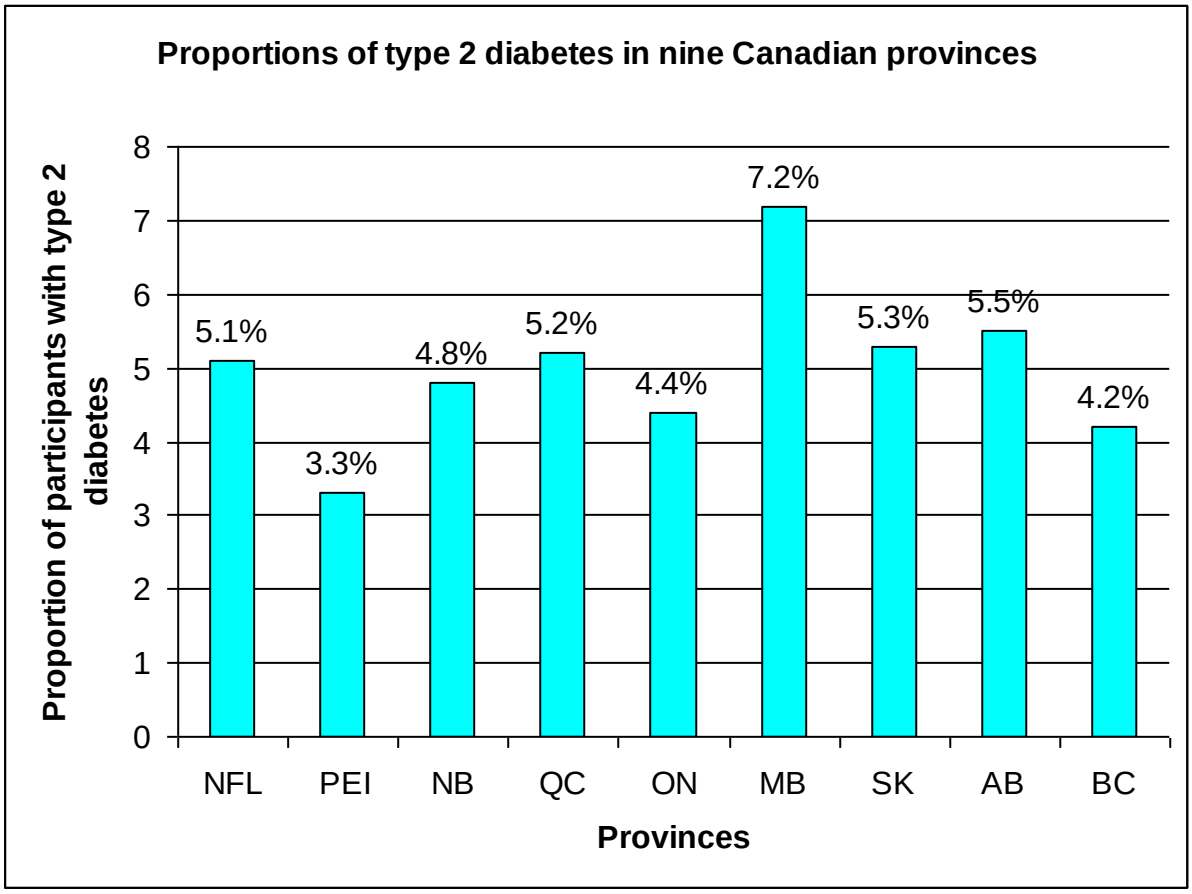


Figure 5.1 Distribution of self-reported, physician-diagnosed type 2 diabetes among provinces from CHHS

5.1.2 Descriptive Analysis

The number of participants and proportions stratified by self-reported, physician-diagnosed type 2 diabetic status for each potential covariate are presented in Table 5.2. The proportions of male and female participants who were diagnosed with self-reported type 2 diabetes were 45.7% and 54.3%, respectively. Based on the self-reported type 2 diabetic status stratified by age group, the proportions of participants with self-reported type 2 diabetes were 27.6%, 26.9% and 45.5% for the age groups 18–44 years, 45–64 years and 65–74 years, respectively. The proportion of participants with self-reported type 2 diabetes was higher among the age groups 45–64 years and 65–74 years compared with the age group for younger participants (18–44 years).

Among the participants with self-reported type 2 diabetes, 38.5% of them lived in rural areas and 61.5% lived in urban areas. Stratifying participants with self-reported type 2 diabetes by employment status, 23.3% of them were full-time workers, 9.4% of them were part-time workers, 23.9% of them were homemakers, 35.5% of them were retired, and 8.0% of them were unemployed. Of participants with self-reported type 2 diabetes, 14.9% of them attended or completed only elementary school, 75.7% of them attended or completed secondary school, and 9.4% of them attended or completed university. The proportion of self-reported type 2 diabetic participants was lowest among the participants with a bachelor's degree or higher education.

Stratifying the self-reported type 2 diabetes status by household income level indicated that about 37.3% of the participants reporting type 2 diabetes were in the household income level $12,000 to $24,999 per year, 31.1% were in the household income level $25,000 to $49,999 per year, 17.2% were in the household income level $12,000 or less per year, followed by 14.4% in the household income level $50,000 or above per year. Among the participants who reported physician-diagnosed type 2 diabetes, 30.6% of them were in the normal weight group (BMI<25 kg/m^2), 31.1% of them were in the overweight group (BMI = 25.0–29.9 kg/m^2), and 32.0% of them were in the obese group (>29.9 kg/m^2). Of participants who had type 2 diabetes, 56.9% of them were involved in physical activity, and 43.1% of them were not involved in physical activity.

Table 5.2 The number of participants in each covariate, stratified by self-reported type 2 diabetic status

	Un-weighted		Weighted	
	Diabetes		Diabetes	
	Yes (%)	No (%)	Yes (%)	No (%)
Sex				
Male	487 (45.7%)	9869 (49.5%)	48.8%	49.5%
Female	579 (54.3%)	10082(50.5%)	51.8%	50.5%
Age Groups				
18–44 years	294 (27.6%)	13275 (66.5%)	29.7%	63.6%
45–64 years	287 (26.9%)	3415 (17.1%)	45.4%	27.1%
65–74 years	485 (45.5%)	3261 (16.4%)	24.9%	9.4%
Location of Residence				
Rural	410 (38.5%)	7340 (36.8%)	21.4%	23.5%
Urban	656 (61.5%)	12611 (63.2%)	78.6%	76.5%
Employment Status				
Retired	378 (35.5%)	2511 (12.6%)	25.6%	9.5%
Part-time	100 (9.4%)	3226 (16.2%)	9.5%	16.9%
Unemployed	85 (8.0%)	1666 (8.4%)	10.8%	8.5%
Homemaker	254 (23.9%)	2815 (14.1%)	21.1%	12.8%
Full-time	248 (23.3%)	9727 (48.8%)	33.0%	52.3%
Education				
Elementary	158 (14.9%)	910 (4.6%)	15.7%	5.6%
Secondary	803 (75.7%)	15750 (79.1%)	73.7%	73.8%
University	100 (9.4%)	3245 (16.3%)	10.7%	20.6%
Household Income				
>$50,000(ref)	133 (14.4%)	4685 (26.4%)	25.3%	35.3%
$25,000–$49,999	287 (31.1%)	6941 (39.0%)	37.4%	37.8%
$12,000–$24,999	344 (37.3%)	4475 (25.2%)	24.7%	17.4%
<$12,000	159 (17.2%)	1679 (9.4%)	12.5%	9.4%
Body Mass Index(BMI)				
BMI<25	278 (30.6%)	8627 (50.4%)	29.5%	52.8%
BMI: 25.0–29.9	340 (37.4%)	5913 (34.5%)	41.3%	34.0%
BMI>29.9	291 (32.0%)	2590 (15.1%)	29.3%	13.2%
Physical Activity				
Yes	606 (56.9%)	12513 (62.7%)	51.5%	62.9%
No	460 (43.1%)	7435 (37.3%)	48.5%	37.1%

5.1.3 Crude prevalence estimation

The Canadian Heart Health Survey (CHHS), a population-based multistage complex survey, was conducted in ten Canadian provinces between 1986 and 1992. Nine Canadian provinces, except Nova Scotia, were included in the present study. Self-reported, physician-diagnosed type 2 diabetes was the outcome variable of interest in this study. A bivariate analysis indicated that the prevalences of self-reported, physician-diagnosed type 2 diabetes in nine Canadian provinces were not identical. The prevalence of self-reported type 2 diabetes in Newfoundland and Labrador (5.4%), New Brunswick (5.5%) and Saskatchewan (5.4%) were higher compared with the other provinces (Table 5.4). Table 5.5 provides the prevalence of self-reported, physician-diagnosed type 2 diabetes stratified by location of residence (rural or urban) in each province. The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among the rural residents in Newfoundland (5.9%), Prince Edward Island (4.6%), Manitoba (6.5%), and Alberta (5.7%) compared with urban residents in the respective provinces. In contrast, the prevalence of self-reported type 2 diabetes was higher among the urban residents in Quebec (5.2%), Ontario (4.3%) and Saskatchewan (5.8%) compared with the rural residents in the same provinces. The overall prevalence of self-reported, physician-diagnosed type 2 diabetes among urban residents (4.7%) was higher than among rural residents (4.2%) (Table 5.3).

The prevalence of self-reported, physician-diagnosed type 2 diabetes stratified by the important covariates is described in Table 5.3. There was no significant difference in the prevalence of self-reported, physician-diagnosed type 2 diabetes between males and females. Participants aged 45 years and above had a higher prevalence of self-reported, physician-diagnosed type 2 diabetes compared with the participants aged less than 45 years. The retired (11.3%), homemaker (7.3%), and unemployed (5.7%) participants had a higher prevalence of self-reported, physician-diagnosed type 2 diabetes compared with the full-time employed (2.9%) participants (Table 5.3).

Participants with only an elementary school education had the highest prevalence (11.7%) of self-reported, physician-diagnosed type 2 diabetes, followed by participants with secondary school education (4.5%). These two groups had higher prevalence of self-reported, physician-diagnosed type 2 diabetes compared with the participants with university education (2.4%). The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among the lower household income (less than $25,000 per year) compared with the participants with household incomes of $49,000 and above per year. The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among the participants who were not involved physical activity (5.9%) compared with the people who were involved in physical activity (3.7%) at least once a week. The prevalence of self-reported, physician-diagnosed type 2 diabetes among obese (BMI>29.9 kg/m^2) participants was the highest, followed by overweight (BMI=25–29.9 kg/m^2) participants. These two groups had a higher prevalence of self-reported, physician-diagnosed type 2 diabetes compared with the normal weight (BMI<25kg/m^2) participants.

Table 5.3 Self-reported type 2 diabetes prevalence (95% C.I.) for all potential covariates included in the model

Covariates	Type 2 diabetes	
	Yes (95% C.I.)	No (95% C.I.)
Sex		
Male	4.5% (3.6–5.8)	95.5% (94.1 – 96.6)
Female	4.6% (3.4–5.9)	95.4% (94.2 – 96.4)
Age group		
18–44 years	2.2% (1.8–2.7)	97.8% (97.3 – 98.3)
45–64 years	7.4% (6.2–8.8)	92.6% (91.2 – 93.8)
65–74 years	11.2% (9.6–13.0)	88.8% (87.0 – 90.4)
Location of Residence		
Rural	4.2% (3.4–5.0)	95.8% (95.0 – 96.6)
Urban	4.7% (4.1–5.2)	95.3% (94.8 – 95.9)
Employment status		
Retired	11.3% (9.3–13.7)	88.7% (86.3 – 90.7)
Part-time	3.2% (2.2–4.5)	96.9% (95.5 – 97.8)
Unemployed	5.7% (3.7–8.7)	94.3% (91.3 – 96.3)
Homemaker	7.3% (5.5–9.6)	92.7% (90.4 – 94.5)
Full-time	2.8% (2.2–3.5)	97.2% (96.5 – 97.8)
Education level		
Elementary	11.7% (8.0–16.8)	88.3% (83.2 – 92.0)
Secondary	4.5% (4.0–5.1)	95.5% (94.9 – 96.0)
University	2.4% (1.6–3.6)	97.6% (96.4 – 98.4)
Household Income		
>$50,000 (ref)	3.2% (2.3–4.5)	96.8% (95.5 – 97.7)
$25,000–$49,999	4.3% (3.6–5.2)	95.7% (94.8 – 96.4)
$12,000–$24,999	6.1% (5.2–7.2)	93.9% (92.8 – 94.8)
<$12,000	5.8% (3.0–10.9)	94.2% (89.1 – 97.1)
Body mass index (BMI)		
BMI<25 kg/m^2	2.7% (2.2–3.4)	97.3% (96.6 – 97.8)
BMI=25–29.9 kg/m^2	5.7% (4.2–7.7)	94.3% (92.3 – 95.8)
BMI>29.9 kg/m^2	10.0% (8.0–12.4)	90.0% (87.6 – 92.0)
Physical activity		
Yes	3.7% (3.2–4.4)	96.3% (95.6 – 96.8)
No	5.9% (4.9–6.9)	94.2% (93.1 – 95.1)

Table 5.4 Diabetes prevalence (%) stratified by type 2 diabetic status for each province

Province	Type 2 diabetes	
	Yes (95% C.I.)	No (95% C.I.)
Newfoundland and Labrador	5.4% (4.8 – 6.2)	94.6% (93.8 – 95.2)
Prince Edward Island	4.1% (3.1 – 5.5)	95.9% (94.5 – 96.9)
New Brunswick	5.5% (4.7 – 6.4)	94.5% (93.6 – 95.3)
Quebec	4.9% (3.9 – 6.2)	95.1% (93.8 – 96.1)
Ontario	4.0% (3.3 – 4.8)	96.0% (95.2 – 96.7)
Manitoba	4.9% (4.1 – 5.8)	95.1% (94.2 – 95.9)
Alberta	4.9% (4.3 – 5.6)	95.1% (94.4 – 95.7)
Saskatchewan	5.4% (3.6 – 8.0)	94.6% (92.0 – 96.4)
British Columbia	4.4% (3.9 – 5.0)	95.6% (95.0 – 96.1)

Table 5.5 Diabetes prevalence (%) stratified by type 2 diabetic status and location of residence for each province

Province	Rural		Urban	
	Type 2 diabetes		Type 2 diabetes	
	Yes (95% C.I.)	No (95% C.I.)	Yes (95% C.I.)	No (95% C.I.)
Newfoundland and Labrador	5.9% (5.1 – 6.8)	94.1% (93.2 – 94.9)	4.8% (3.7 – 6.2)	95.2% (93.8 – 96.3)
Prince Edward Island	4.6% (3.1 – 6.7)	95.4% (93.3 – 96.9)	3.5% (2.4 – 5.2)	96.5% (94.8 – 97.6)
New Brunswick	5.5% (4.3 – 7.0)	94.5% (93.0 – 95.7)	5.4% (3.9 – 7.3)	94.6% (92.7 – 96.1)
Quebec	3.6% (1.9 – 6.6)	96.4% (93.4 – 98.1)	5.2% (4.1 – 6.5)	94.8% (93.5 – 95.9)
Ontario	3.0% (2.0 – 4.4)	97.0% (95.6 – 98.0)	4.3% (2.8 – 6.5)	95.7% (93.5 – 97.20
Manitoba	6.5% (4.7 –8.9)	93.5% (91.1 – 95.3)	4.4% (3.3 – 5.7)	95.6% (94.3 – 96.7)
Alberta	5.7% (4.7 – 6.9)	94.3% (93.1 – 95.3)	4.6% (3.4 – 6.2)	95.4% (93.8 – 96.6)
Saskatchewan	4.5% (4.0 – 5.1)	95.5% (94.9 – 96.0)	5.8% (4.5 – 7.5)	94.2% (92.5 – 95.5)
British Columbia	4.7% (3.5 – 6.3)	95.3% (93.7 –96.5)	4.3% (3.3 – 5.7)	95.7% (94.5 – 96.7)

5.1.4 Modeling approach for cross-sectional complex survey data and results

In order to compare the MM-SW technique and the SR-RV estimation technique, the estimated regression coefficients and the standard errors obtained from the two statistical techniques were estimated. The MM-SW technique and the SR-RV estimation technique were used to analyze the Canadian Heart Health Survey (CHHS) data. Both statistical techniques take into account the complexities of complex surveys, but the way of accounting for these complexities are different. In the SR-RV estimation technique, the pseudo maximum likelihood (PML) function was used to estimate the regression coefficients. Bootstrap re-sampling methods were used to estimate the standard errors of the regression coefficients. Five hundred bootstrap weights, including the final weight, which were available in the CHHS, were used to estimate the standard errors of the regression coefficients using a bootstrap re-sampling method.

In the MM-SW technique, the two-level random-intercept logistic regression models were used to analyze the CHHS data sets. The individuals represented the level 1 unit, and the primary sampling units (PSU) represented the level 2 units in the CHHS dataset. Multilevel pseudo-maximum likelihood was used to estimate the regression coefficients via adaptive quadrature with scaled weights in the multilevel modeling technique. Appropriate scaling of level 1 weight might reduce the bias of the standard errors [7, 57, 58, 60]. The standard errors of the regression coefficients were estimated using the sandwich estimator, which takes into account the stratification and the clustering, the two important characteristics of complex survey data. The statistical analysis based on multilevel modeling was conducted using "GLLAMM" in STATA software.

The outcome variable of interest was self-reported, physician-diagnosed type 2 diabetes, which was dichotomous (yes, no), where "yes" means participants who had type 2 diabetes and "no" means those who didn't have type 2 diabetes. The independent covariates, considered to be risk factors for the prevalence of type 2 diabetes, were selected using standard model-building techniques. A

bivariate analysis was conducted with the outcome variable of self-reported type 2 diabetes (yes, no) and important covariates thought to be risk factors for the prevalence of type 2 diabetes. Those covariates with $p \leq 0.25$ or with biological significance were selected for the final model. The selected covariates for the final model were as follows: sex, age group, location of residence (rural or urban), household income per year, employment status, physical activity, and body mass index. All of the covariates that were included in the final model were categorical. The estimated regression coefficients and the standard errors based on the analyses of the CHHS using the MM-SW technique are presented in Table 5.6. Multilevel pseudo-maximum likelihood (MPML) was used to estimate the regression coefficient estimators, and sandwich estimators were used to estimate the standard errors of the regression coefficient estimators after taking into account the design effects of complex surveys, such as stratification and clustering. The regression coefficient estimators and their standard errors based on the standard regression–robust variance estimation technique are also presented in Table 5.6.

To our knowledge, no goodness-of-fit test for survey data is available for the multilevel modeling technique. This could be an active research area for future research. In the SR-RV estimation technique, pseudo maximum likelihood was used to estimate the regression coefficient estimators, and bootstrap methods were used to estimate the standard errors. After fitting the logistic regression model using the standard regression-robust variance estimation technique, a goodness-of-fit test was used to see whether the model fit the survey data adequately or not. The command *estat gof* in STATA applied the residual goodness-of-fit test for the survey data. This goodness-of-fit test for survey data was discussed in section 3.1.3.1. The goodness-of-fit test indicated that the final logistic regression model fit the survey data with p-value 0.105 Therefore; the final logistic regression model fit the survey data adequately.

Finally, the estimated regression coefficients and their standard errors from the multilevel modeling (random-intercept logistic regression)–scaled weights technique and the standard regression (logistic regression)–robust variance estimation technique were different. The standard errors of the regression coefficients were higher for the SR-RV estimation technique compared with the MM-SW technique. The estimated 95% confidence intervals for the regression coefficients were wider for the SR-RV estimation technique compared with the MM-SW technique.

5.1.5 Comparison between the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique based on the analysis of CHHS

The first objective of the thesis was to compare the MM-SW technique and the SR-RV estimation technique based on analyses of cross-sectional complex survey data. These two statistical techniques were applied to analyze the Canadian Heart Health Survey (CHHS) data. The estimated regression coefficients, standard errors and the 95% confidence intervals obtained from both statistical techniques was presented in Table 5.6. The results, based on the analyses of the CHHS, indicated that the estimated regression coefficients were not similar between the two statistical techniques. The MM-SW technique used multilevel pseudo maximum likelihood (MPML) to estimate the regression coefficients, and the SR-RV technique used pseudo maximum likelihood (PML) to estimate the regression coefficients. The probability weights of sampling units for each level of complex survey data were used in MPML. In contrast, only the overall probability weights of level 1 units were used in PML. The scaling of probability weights for level 1 unit was used in MPML, whereas raw probability weights were used in PML. These are possible reasons for the differences in regression coefficients and the standard errors between these two statistical techniques.

The standard errors of each estimated regression coefficient based on the MM-SW technique were smaller than the standard errors of regression coefficients obtained from the SR-RV estimation

technique. Sandwich variance estimators were used to estimate standard errors after taking into account the effects of sampling design in the MM-SW technique, while the bootstrap re-sampling technique was used to estimate the standard errors of the regression coefficients in the SR-RV estimation technique. These are possible reasons for the different standard errors obtained from the two statistical techniques. The sandwich variance estimator may underestimate the variance of parameters based on design-based analysis [36].

The 95% confidence intervals (95% C.I.) for the estimated regression coefficients are narrower for the MM-SW technique than SR-RV estimation technique. These results indicate that the performance of the MM-SW technique might be better than the SR-RV estimation technique for analyzing cross-sectional complex survey data. The results based on the analysis of a single real life complex survey data may not possible to generalize. A Monte Carlo simulation study based on cross-sectional complex survey data may provide firm results to compare the performance of these statistical techniques which is the third objective of the thesis.

Table 5.6 Parameter estimates (standard errors) and their 95% confidence intervals based on the CHHS

	Multilevel modeling		Standard regression	
	Scaled weight		Robust(Bootstrap)	
Covariates	Estimates (SE)	95% C.I.	Estimates (SE)	95% C.I.
Intercepts	–4.72 (0.26)*	–5.23, –4.21	–4.89 (0.28)*	–5.44, –4.34
Age				
18–44 years (ref)				
45–64 years	0.24 (0.19)	–0.13, 0.61	0.25 (0.21)	–0.16, 0.66
65–74 years	0.49 (0.18)*	0.14, 0.84	0.74 (0.33)*	0.10, 1.37
Location of residence				
Rural (ref)				
Urban	0.09 (0.12)	–0.15, 0.32	0.20 (0.15)	–0.08, 0.49
Sex				
Female (ref)				
Male	–0.60 (0.19)*	–0.98, –0.23	–0.41 (0.34)	–1.08, 0.26
Education				
University (ref)				
Secondary	0.34 (0.20)	–0.05, 0.72	0.40 (0.26)	–0.11, 0.91
Elementary	0.50 (0.33)	–0.15, 1.14	0.69 (0.34)*	0.02, 1.37
Household Income				
>$50,000 (ref)				
$25,000–$49,999	0.22 (0.16)	–0.91, 0.54	0.17 (0.32)	–0.47, 0.80
$12,000–$24,999	0.43 (0.17)*	0.10, 0.76	0.17 (0.23)	–0.28, 0.63
<$12,000	0.41 (0.20)*	0.02,0.80	–0.04 (0.24)	–0.50, 0.43
Employment Status				
Full-time (ref)				
Part-time /students	–0.05 (0.23)	–0.51, 0.41	–0.17 (0.24)	–0.57, 0.26
Unemployment	0.85 (0.24)*	0.38, 1.32	0.60 (0.30)*	0.01, 1.18
Homemaker	0.80 (0.18)*	0.45, 1.16	0.69 (0.27)*	0.15, 1.23
Retired	0.83 (0.21)*	0.42, 1.24	0.59 (0.36)	–0.12,1.30
Physical Activity				
Yes (ref)				
No	0.27 (0.12)*	0.03, 0.52	0.36 (0.20)	–0.03, 0.75
Body Mass Index (BMI)				
BMI: <25(ref)				
BMI: 25.0–29.9	0.41 (0.17)*	0.06, 0.75	0.60 (0.23)*	0.15, 1.05
BMI: >29.9	1.15 (0.13)*	0.90, 1.41	1.24 (0.16)*	0.93, 1.56

Cont'd Table 5.6

	Multilevel modeling		Standard regression	
	Scaled weight		Robust(Bootstrap)	
Interaction				
Age groups * sex				
18–44yrs*female (ref)				
45–64yrs*male	1.26 (0.26)*	0.76, 1.77	1.21 (0.38)*	0.47, 1.95
65–74yrs*male	0.87 (0.28)*	0.32, 1.42	0.84 (0.35)*	0.16, 1.52

* indicates p-value≤0.05

5.2 Models for longitudinal complex survey data

The second objective of this thesis was to compare the MM-SW technique and the SR-RV estimation technique based on analyses of longitudinal complex survey data. In order to compare these two statistical techniques, they were utilized to analyze the longitudinal national population health survey (NPHS) data. Repeated measurements of each subject over time were an additional character of longitudinal data. Statistical analyses of longitudinal complex survey data is more complicated compared with cross-sectional complex survey data because of repeated within-subject measurements [1, 25, and 48]. The chosen statistical technique should take into account within-subject correlation due to repeated measurements, including complex survey design effects such as stratification and clustering in longitudinal complex survey data.

Sections 5.2.1 and 5.2.2 describe the characteristics of the study population and the descriptive analysis, respectively. The results related to crude prevalence estimation of self-reported, physician-diagnosed type 2 diabetes are presented in section 5.2.3. The results based on the two multi-variable techniques of interest are discussed in section 5.2.4. The results based on the Monte Carlo simulation study are presented in section 5.3. Section 5.4 describes the interpretation of results obtained from the analysis of the CHHS and NPHS datasets.

5.2.1 Study Population

The total number of participants in the NPHS from the ten Canadian provinces was 17,276 in 1994–95 (Cycle 1). All participants (14,117) who were 18 years and older at the beginning of cycle 1 (1994–95) were included in our analysis. People living on Indian Reserves and Crown lands, residents of health institutions, full-time members of the Canadian Forces Bases and those living in remote areas in Ontario and Quebec were excluded from this survey. The number of people who

provided a full response to all six cycles is shown in Table 4.2. Detail descriptions of all participants who were included in the NPHS were given in Chapter 4.

5.2.2 Descriptive analysis

The number of participants from all Canadian provinces who were included in this study was 14,117 in Cycle 1 (1994–95); they were followed up every two years until 2005. The numbers of participants, stratified by cycles and self-reported, physician-diagnosed type 2 diabetic status, are presented in Table 5.7. The numbers of self-reported, physician-diagnosed type 2 diabetic cases increased over time from Cycle 1 to Cycle 6. The number of non-diabetic participants decreased over time.

Table 5.7 Distribution of self-reported, physician-diagnosed type 2 diabetic and non-diabetic participants (%), stratified by cycles

Cycles	Diabetes		
	Yes (%)	No (%)	Totals
Cycle 1 (1994–95)	529 (14.7%)	13565 (19.4%)	14094
Cycle 2 (1996–97)	557 (15.5%)	12693 (18.2%)	13250
Cycle 3 (1998–99)	567 (15.8%)	12006 (17.2%)	12573
Cycle 4 (2000–01)	600 (16.7%)	11262 (16.1%)	11862
Cycle 5 (2002–03)	661 (18.4%)	10,521 (15.0%)	11182
Cycle 6 (2004–05)	681 (18.9%)	9890 (14.1%)	10571

The distribution of self-reported, physician-diagnosed type 2 diabetic and non-diabetic cases based on Cycle 1 (1994–95), according to potential covariates, is shown in Table 5.8. The percentage of self-reported, physician-diagnosed type 2 diabetic cases was higher among females (55.4%) compared with males (44.6%). The percentage of self-reported, physician-diagnosed type 2 diabetic cases, stratified by age group is as follows: 18–44 years, 45–65 years, and >65 years were 13.8%, 32.3%, and 53.9%, respectively. The percentage of self-reported, physician-diagnosed type 2 diabetic cases was higher among urban (59.2%) residences compared with rural (40.8%)

residences in Canada. The self-reported, physician-diagnosed type 2 diabetic status was stratified by education levels. The percentage of self-reported type 2 diabetic cases was much higher among participants with elementary (53.6%) and secondary (33.3%) school levels of education compared with participants with university degrees (13.1%).

The percentages of self-reported, physician-diagnosed type 2 diabetic cases for the following ranges of household incomes: <$15,000, $15,000–$29,999, $30,000–$49,999, and >$50,000 were 30.7%, 34.5%, 21.8% and 13.1%, respectively. The results based on the descriptive analysis indicated that the percentage of self-reported, physician-diagnosed type 2 diabetes was higher among participants with lower household income.

Based on the stratification of self-reported, physician-diagnosed type 2 diabetic cases by the status of physical exercise, the percentage of participants who were not involved in any physical exercise was 39.1% and the percentage of participants who were occasionally or regularly involved in physical exercise was 61%.

The participants were stratified by self-reported, physician-diagnosed type 2 diabetic status and body mass index levels: normal weight (<25 kg/m^2), overweight (25–29.9kg/m^2), and obese (>29.9 kg/m^2). The percentage of self-reported, physician-diagnosed type 2 diabetic cases was higher among participants who were overweight (39.2%) or obese (31.9%) compared with participants with a normal weight (28.9%).

Self-reported, physician-diagnosed type 2 diabetic participants were stratified based on their birth parents' type 2 diabetic status. The percentage (31.2%) of self-reported, physician-diagnosed type 2 diabetic participants whose birth mother had type 2 diabetes was higher compared with participants (12.1%) whose birth mother did not have type 2 diabetes.

The percentage (19.6%) of self-reported, physician-diagnosed type 2 diabetic participants whose birth father had type 2 diabetes was higher compared with participants (10.1%) whose birth

father did not have type 2 diabetes. The distribution of participants in the ten Canadian provinces over the six cycles is presented in Table 5.9. The participants' rate of diabetes was highest in Ontario (25.3%), followed by Quebec (17.1%), British Columbia (10.1%) and Alberta (8.8%). The participants' rate of diabetes was lowest in Newfoundland and Labrador (6.0%).

Table 5.8 Number (%) of self-reported, physician-diagnosed type 2 diabetic cases according to the potential risk factors based on Cycle 1(1994-95)

Covariates	Diabetes		
	Yes (%)	No (%)	Total
Sex			
Male	236 (44.6%)	6,210 (45.8%)	6446
Female	293 (55.4%)	7,355 (54.2%)	7648
Age group			
18–44 years	73 (13.8%)	7,405 (54.6%)	7478
45–65 years	171 (32.3%)	3,711 (27.3%)	3882
>65 years	285 (53.9%)	2,449 (18.1%)	2734
Location of residence			
Rural	211 (40.8%)	4524 (34.2%)	4735
Urban	306 (59.2%)	8689 (65.8%)	8995
Level of education			
University	69 (13.1%)	3583 (26.5%)	3652
Secondary	175 (33.3%)	6065 (44.8%)	6240
elementary	282 (53.6%)	3888 (28.7%)	4170
Household income per year			
>$50,000	66 (13.1%)	3733 (28.8%)	3799
$30,000–$49,999	110 (21.8%)	3541 (27.3%)	3651
$15,000–$29,999	174 (34.5%)	3268 (25.2%)	3442
<$15,000	155 (30.7%)	2413 (18.6%)	2568
Physical Activity			
Regular	295 (60.99%)	9439 (74.7%)	9734
Infrequent	189 (39.1%)	3191 (25.3%)	3380
Body mass index (BMI)			
Normal weight (BMI<25kg/m^2)	146 (28.9%)	6594 (51.3%)	6740
Overweight (BMI=25.0 – 29.9kg/m^2)	198 (39.2%)	4551 (35.3%)	4749
Obese (BMI>29.9kg/m^2)	161 (31.9%)	1703 (13.3%)	1864
Mother had diabetes			
Yes	97 (31.2%)	1235 (12.1%)	1332
No	214 (68.8%)	8961 (87.9%)	9175
Father had diabetes			
Yes	60 (19.6%)	1004 (10.1%)	1064
No	246 (80.4%)	8986 (89.9%)	9232

Table 5.9 Distribution of participants (%) stratified by cycles and provinces

Province	Cycle 1	Cycle 2	Cycle 3	Cycle 4	Cycle 5	Cycle 6
Newfoundland and Labrador	849 (6.0%)	841 (5.7%)	831 (5.5%)	839 (5.4%)	861 (5.5%)	864 (5.4%)
Prince Edward Island	868 (6.2%)	867 (5.9%)	867 (5.8%)	878 (5.7%)	885 (5.6%)	893 (5.6%)
Novas Scotia	895 (6.3%)	919 (6.3%)	929 (6.2%)	944 (6.1%)	959 (6.1%)	985 (6.2%)
New Brunswick	916 (6.5%)	939 (6.4%)	974 (6.5%)	993 (6.4%)	1005 (6.4/%)	1017 (6.4%)
Quebec	2417 (17.1%)	2522 (17.2%)	2600 (17.2%)	2667 (17.3%)	2746 (17.5%)	2802 (17.5%)
Manitoba	985 (7.0%)	1016 (6.9%)	1024 (6.8%)	1040 (6.7%)	1054 (6.7%)	1067 (6.7%)
Alberta	1236 (8.8%)	1330 (9.1%)	1423 (9.4%)	1484 (9.6%)	1541 (9.8%)	1592 (9.9%)
Saskatchewan	955 (6.8%)	965 (6.6%)	986 (6.5%)	984 (6.4%)	995 (6.3%)	1005 (6.3%)
British Columbia	1428 (10.1%)	1532 (10.5%)	1587 (10.5%)	1617 (10.5%)	1620 (10.3%)	1652 (10.3%)
Ontario	3568 (25.3%)	3720 (25.4%)	3867 (25.6%)	3991 (25.9%)	4041 (25.1%)	4146 (25.9%)

Table 5.10 describes the prevalence of self-reported, physician-diagnosed type 2 diabetes among Canadian adults (18 years or older), stratified by the cycles. The prevalence of self-reported type 2 diabetes and the 95% confidence intervals (95% C.I.) were determined using the BOOTVAR macro provided by Statistics Canada. The trend of prevalence of self-reported, physician-diagnosed type 2 diabetes indicates that the numbers of self-reported, physician-diagnosed type 2 diabetic cases are increasing among Canadians participants.

Table 5.10 Prevalence of type 2 diabetes (95% confidence interval) stratified by cycles

Cycle	Prevalence	95% Confidence Interval
Cycle 1 (1994–1995)	3.4	3.0 – 3.8
Cycle 2 (1996–1997)	3.8	3.4 – 4.2
Cycle 3 (1998–1999)	3.9	3.5 – 4.3
Cycle 4 (2000–2001)	4.5	4.3 – 4.7
Cycle 5 (2002–2003)	5.1	4.6 – 5.7
Cycle 6 (2004–2005)	5.7	5.1 – 6.2

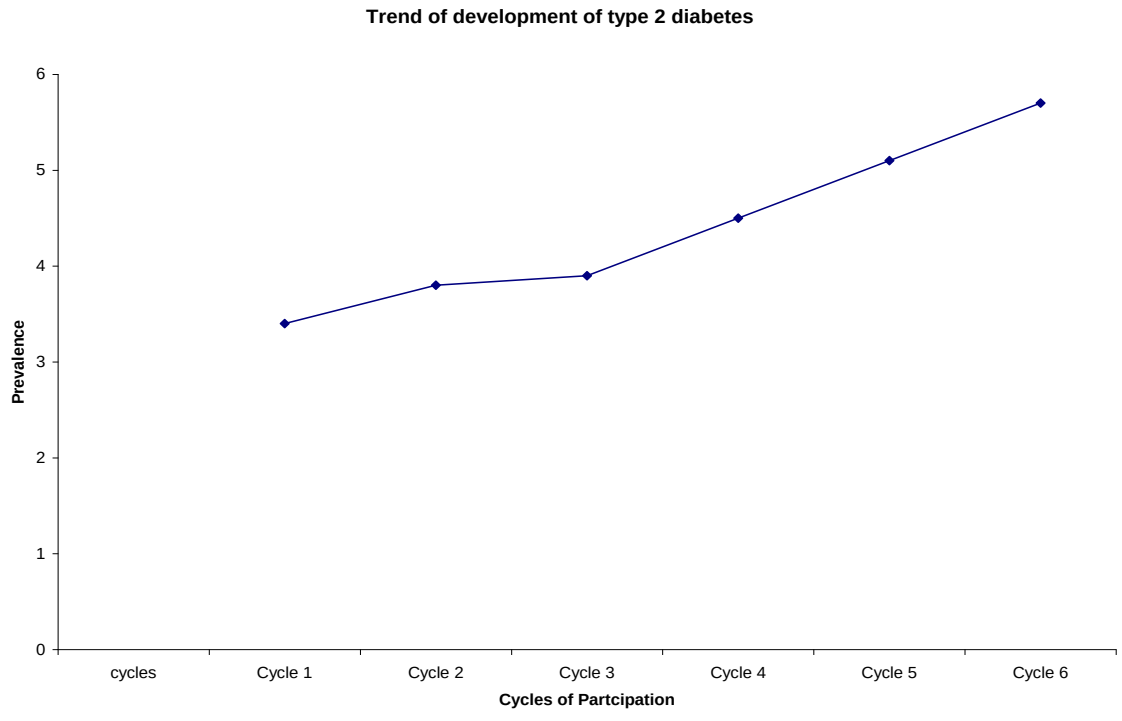


Figure 5.2 Prevalence of self-reported, physician-diagnosed type 2 diabetes over time

5.2.3 Estimation of crude prevalence

The prevalence of self-reported, physician-diagnosed type 2 diabetes stratified by the potential covariates that were included in the final multivariable logistic regression model is described in Table 5.11. The BOOTVAR macro provided by Statistics Canada was used to calculate the prevalence and their 95% confidence intervals.

The prevalence of self-reported, physician-diagnosed type 2 diabetes increased with age. The prevalence of self-reported, physician-diagnosed type 2 diabetes among participants in age group 44–64 years and age > 65 years was higher compared with the participants in age group 18–44 years. The prevalence of self-reported, physician-diagnosed type 2 diabetes had an increasing trend over time among participants aged 45–64 years and 65 years and above. The prevalence of self-reported, physician-diagnosed type 2 diabetes was slightly higher among rural residents compared with urban residents in Canada, but this did not change over time. The male participants had a slightly higher prevalence of self-reported, physician-diagnosed type 2 diabetes compared with females. The prevalence of self-reported, physician-diagnosed type 2 diabetes among males increased from Cycle 1 to Cycle 3 and then slightly decreased and became steady from Cycle 4 to Cycle 6. It was almost unchanged among females over time. Participants with only an elementary school education had the highest prevalence of self-reported, physician-diagnosed type 2 diabetes, followed by participants with secondary school education. The prevalence of self-reported type 2 diabetes increased rapidly over time among both groups of participants compared with participants with a university degree. The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among participants with lower education. The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among participants with irregular or infrequent physical exercise habits compared with participants with regular physical exercise habits. The prevalence of

self-reported, physician-diagnosed type 2 diabetes increased rapidly among participants with irregular or infrequent physical exercise habits, but slowly among people with regular physical exercise habits over time. The prevalence of self-reported, physician-diagnosed type 2 diabetes was highest among obese (BMI> 29.9 kg/m^2) participants compared with participants with a normal (BMI<25 kg/m^2) weight. The prevalence of self-reported, physician-diagnosed type 2 diabetes was also higher among overweight participants than among participants with normal weight.
The rate of prevalence of self-reported, physician-diagnosed type 2 diabetes among obese and overweight participants also increased over time compared with normal weight participants.

The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among participants whose birth mother had type 2 diabetes than among participants whose birth mother did not have type 2 diabetes, and also the rate of prevalence over time increased among participants whose birth mother had type 2 diabetes compared with participants whose birth mother did not have type 2 diabetes. The prevalence of self-reported, physician-diagnosed type 2 diabetes was higher among participants whose birth father had type 2 diabetes than among participants whose birth father did not have type 2 diabetes. The prevalence of self-reported, physician-diagnosed type 2 diabetes increased over time among participants whose birth father had type 2 diabetes compared with the participants whose birth father did not have type 2 diabetes.

Table 5.11 Self-reported and physician diagnosed type 2 diabetes prevalence (95% confidence interval) for potential covariates included in the final multivariable logistic regression model

Covariate	Cycle 1 (1994–95)	Cycle 2 (1996–97)	Cycle 3 (1998–99)	Cycle 4 (2000–01)	Cycle 5 (2002–03)	Cycle 6 (2004–05)
Age group						
18–44 years	1.0 (0.7 – 1.2)	0.9 (0.6 – 1.2)	0.9 (0.1 – 0.7)	1.0 (0.7 – 1.4)	0.9 (0.6 – 1.2)	1.3 (0.9 – 1.7)
45–64 years	4.0 (3.1 – 4.9)	4.8 (3.9 – 5.7)	4.7 (3.9 – 5.6)	5.6 (4.6 – 6.6)	6.6 (5.6 – 7.7)	6.1 (5.1 – 7.1)
65–75 years	11.2 (9.4 – 13.0)	11.6 (10.0 – 13.3)	11.7 (10.2 –13.2)	12.7 (10.9 – 14.6)	14.0 (12.3 – 15.8)	15.3 (13.4 – 17.3)
Area						
Rural	3.9 (3.1 – 4.7)	4.2 (3.5 – 5.0)	4.7 (3.6 – 5.8)	3.8 (3.0 – 4.7)	4.4 (3.5 – 5.3)	4.3 (3.2 – 5.4)
Urban	3.2 (2.8 – 3.7)	3.6 (3.1 – 4.1)	3.8 (3.3 – 4.2)	3.4 (3.0 – 3.9)	3.6 (3.2 – 4.0)	3.7 (3.3 – 4.1)
Sex						
Male	3.5 (2.9 – 4.0)	4.0 (3.3 – 4.6)	4.3 (3.6 – 4.9)	3.7 (3.1 – 4.3)	3.8 (3.2 – 4.4)	3.9 (3.3 – 4.5)
Female	3.3 (2.7 – 3.8)	3.6 (3.0 – 4.1)	3.5 (3.0 – 4.0)	3.7 (2.8 – 3.7)	3.7 (3.2 – 4.1)	3.6 (3.2 – 4.1)
Education						
Elementary	6.1 (5.1 – 7.1)	7.2 (6.0 – 8.3)	7.1 (6.0 – 8.1)	8.2 (6.8 – 9.7)	9.8 (8.2 – 11.4)	10.4 (8.7 – 12.0)
Secondary	2.7 (2.2 – 3.2)	2.9 (2.4 – 3.5)	3.3 (2.8 – 3.9)	4.1 (3.4 – 4.7)	4.9 (4.2 – 5.6)	5.7 (4.9 – 6.4)
Bachelor and above	1.8 (1.2 – 2.4)	2.2 (1.5 – 2.9)	2.1 (1.4 – 2.7)	2.5 (1.8 – 3.1)	2.8 (2.2 – 3.5)	3.2 (2.5 – 3.9)

Cont'd Table 5.11

Covariates	Cycle 1 (1994–95)	Cycle 2 (1996–97)	Cycle 3 (1998–99)	Cycle 4 (2000–01)	Cycle 5 (2002–03)	Cycle 6 (2004–05)
			Physical			
Exercise						
Infrequent	4.8 (3.9 – 5.7)	6.0 (4.9 – 7.0)	6.0 (4.8 – 7.2)	6.7 (5.4 – 8.0)	8.5 (6.8 – 10.2)	9.8 (8.0 – 11.6)
Regular	2.9 (2.4 – 3.3)	3.0 (2.6 – 3.5)	3.4 (2.9 – 3.9)	3.7 (3.3 – 4.2)	4.3 (3.8 – 4.8)	4.7 (4.1 – 5.3)
Body mass index(BMI)						
Normal Weight	2.1 (1.6 – 2.5)	2.2 (1.7 – 2.7)	2.1 (1.7 – 2.5)	2.4 (1.8 – 2.9)	2.5 (2.0 – 3.0)	3.2 (2.4 – 3.9)
Overweight	4.0 (3.2 – 4.8)	4.6 (3.9 – 5.4)	4.4 (3.7 – 5.2)	5.5 (4.5 – 6.4)	5.8 (4.8 – 6.8)	5.7 (4.8 – 6.6)
Obese	7.5 (6.0 – 9.0)	7.9 (6.4 – 9.4)	8.3 (6.6 – 9.9)	8.2 (6.8 – 9.6)	10.6 (8.9 – 12.3)	11.4 (9.7 – 13.1)
Mother had diabetes						
Yes	7.0 (5.2 – 8.6)	7.4 (5.7 – 9.0)	9.0 (7.1 – 10.9)	9.8 (7.8 – 11.8)	11.4 (9.2 – 13.6)	10.8 (8.7 – 12.9)
No	2.1 (1.7 – 2.6)	2.5 (2.1 – 2.9)	2.9 (2.5 – 3.3)	3.3 (2.8 – 3.8)	3.6 (3.2 – 4.1)	4.0 (3.5 – 4.5)
Father had Diabetes						
Yes	5.3 (3.5 – 7.1)	6.1 (4.2 – 8.0)	6.6 (4.7 – 8.4	6.5 (4.6 – 8.4)	8.0 (5.9 – 10.0)	8.7 (6.5 –10.8)
No	2.5 (2.1 – 2.9)	2.8 (2.4 – 3.3)	3.3 (2.9 – 3.8)	3.8 (3.3 – 4.2)	4.2 (3.8 – 4.7)	4.4 (3.9 – 4.9)

5.2.4 Modeling approach for longitudinal complex survey data

The second objective of this thesis was to compare the MM-SW technique and the SR-RV estimation technique through an analysis of longitudinal complex survey data.

For the SR-RV estimation technique, the marginal logistic regression model was utilized to analyze the National Population Health Survey (NPHS) data. The dichotomous outcome variable of interest was self-reported, physician-diagnosed type 2 diabetes with responses of 'yes' or 'no'. The response for the outcome variable was based on the question, "Do you have any of the following long-term conditions that have been diagnosed by a health professional? – Diabetes?" The independent covariates for the multivariable model were selected based on the standard model building technique. A bivariate analysis was conducted with the outcome variable of self-reported, physician-diagnosed type 2 diabetes (yes, no) and selected covariates that are thought to be risk factors for type 2 diabetes. Covariates with $p \leq 0.25$ or biological significance were included in the multivariable model, and covariates with $p \leq 0.05$ were retained in the final model. The selected covariates for the final models were: sex, age group, location of residence (rural or urban), education level, household income per year, physical activity, body mass index, and birth mother had diabetes, and birth father had diabetes. All covariates that were retained in the final model were categorical. All statistical analyses were conducted based on weighted data. The generalized estimating equations (GEE) based on quasi-likelihood function were used to estimate the regression coefficients and the adjusted odds ratios using GENMOD in the SAS® software program.

After selecting all significant covariates and interactions for final model, we conducted a goodness-of-fit test to determine whether or not the model was a good fit to the observed data. Since the GEE model was based on the quasi-likelihood method, the QIC and $QIC_u(R)$ statistics were used to test the adequacy of the model fitting. These were discussed in section 3.2.2.6. The final model was selected based on the smallest difference between the QIC and $QIC_u(R)$ values [SAS

documents]. The GEE model with three correlation structures: Autoregressive (AR(1)), exchangeable (EXCH), and unstructured (UN) were performed. A convergence problem occurred with the unstructured correlation for the final model. Generally, a convergence problem occurs when the Hessian matrix is not definitely positive. I selected the exchangeable within-subject correlation structure for the final GEE model, based on smallest values of the QIC.

The BOOTVAR program provided by Statistics Canada contains a set of macros that was available along with the NPHS data. The BOOTVAR macro takes into account the design features (stratification, clustering and unequal probability of selection) for longitudinal complex survey data in order to estimate valid standard errors of regression coefficient estimators. Five hundred sets of bootstrap weights based on the re-sampling technique were used to estimate the standard errors of the regression coefficients estimators and the corresponding 95% confidence intervals (95% C.I.) of the regression coefficients estimators and odds ratios. The regression coefficients estimators, their standard errors and the 95% confidence intervals based on the SR-RV estimation technique are presented in Table 5.12.

In the MM-SW technique, two-level random-intercept logistic regression models were used to analyze the NPHS data. In this technique, level 1 represented the repeated measurements within-subject, and subjects indicated level 2 units. The outcome variable and the independent covariates of interest were exactly the same as in the SR-RV estimation technique. In order to select the independent covariates for the initial multivariable model, the standard model building technique was used, based on a selection criteria of $p \leq 0.25$, which means the independent covariates with $p \leq 0.25$ in the univariate analysis were selected for the initial multivariable model.

In the multilevel modeling technique for longitudinal complex survey data, repeated measurements were nested within subjects and subjects were nested within the PSU (primary sampling unit). The "GLLAMM" procedure in the STATA software program was used to fit the data in random-

intercept logistic regression models. In the "GLLAMM" procedure, multilevel pseudo maximum likelihood was used via ordinary quadrature to estimate the regression coefficient estimators and odds ratios with scaled weights. The scaling method of weights was discussed in Chapter 3. The sandwich estimator method was used to determine the valid standard errors of the regression coefficients. The regression coefficients estimators, their standard errors and 95% confidence intervals (95% C.I.) based on the MM-SW technique using the NPHS are shown in Table 5.12.

5.2.5 Risk factors for type 2 diabetes based on the NPHS

Based on the MM-SW technique, the significant predictors of diabetes were age group (18–44 years, 45–64 years and 65 years and above), education level (elementary, secondary and university), household income (<$12,000, $12,000–$24,999, $25,000–$49,999 , and >$50,000), body mass index (BMI<25 kg/m^2, BMI = 25.0–29.9 kg/m^2, BMI>29.9 kg/m^2), mother had type 2 diabetes (yes, no), father had type 2 diabetes (yes, no) and cycles (time1, time2, time3, time4, time5 and time6). No interaction terms were significant at $p \leq 0.05$.

Based on the SR-RV estimation technique, the significant predictors of diabetes at the $p \leq 0.05$ level were age (18–44 years, 45–64 years and 65 years and above), sex (male, female), education level (elementary, secondary and university), household income (<$12,000, $12,000–$24,999, $25,000–$49,999, and >$50,000), body mass index (BMI<25 kg/m^2, BMI = 25.0–29.9 kg/m^2, BMI>29.9 kg/m^2), mother had type 2 diabetes (yes, no), father had type 2 diabetes(yes, no), cycles (time1, time2, time3, time4, time5 and time6), and an interaction term—sex*household income. In the SR-RV estimation technique, the interaction between sex and household income was significant at $p \leq 0.05$, whereas no interaction was significant at the $p \leq 0.05$ significance level in the MM-SW technique.

The results based on the analyses of the NPHS data indicated that the regression coefficient estimators and their standard errors for all covariates in the MM-SW technique were larger compared with the SR-RV estimation technique (Table 5.12).

5.2.6 Comparison of the results obtained from the two techniques

The results, based on the analysis of the NPHS, indicated that the estimated regression coefficients were not similar between the two techniques. The estimated regression coefficient estimators were higher for the MM-SW technique compared with the SR-RV estimation technique which was expected. The MM-SW technique produced higher standard errors of the regression coefficient estimators compared with the standard errors of estimated regression coefficients in the SR-RV estimation technique. Consequently, the 95% confidence intervals (95% C.I.) for the estimated regression coefficient estimators were narrower in the SR-RV estimation technique. Both statistical techniques provided the same number of significant predictors associated with the prevalence of self-reported, physician-diagnosed type 2 diabetes. The common significant predictors for both models were: age group, education level, household income, body mass index, mother had diabetes, father had diabetes, and time of observation. The variable sex was not significant in multilevel model, the interaction term household income and sex was significant in regression model but not in multilevel model. The interaction between sex and income level was significant ($p \leq 0.05$) in the SR-RV estimation technique but was not significant in the MM-SW technique. There are many possible reasons for this difference. The estimated regression coefficients obtained from multilevel modeling and standard regression (generalized estimating equations) can be approximately connected by the following relationship:

$\beta_{sr} \approx \frac{\beta_{ml}}{\sqrt{1+0.3\sigma^2}}$ where $Var(\zeta_j) = \sigma^2$, β_{sr} and β_{ml} denote the regression coefficients obtained from SR-RV estimation technique and MM-SW respectively. From this relationship, it is clear that if $\sigma^2 = 0$ then the regression coefficients obtained from both methods are equal, and if $\sigma^2 > 0$ then $|\beta_{sr}| < |\beta_{ml}|$ that means the regression coefficients obtained from multilevel modeling are larger than the regression coefficients obtained from standard regression. For example, the regression coefficients (β_{ml}) of age(45-64yrs) is 2.29 obtained from multilevel modeling-scaled weights technique (Table 5.12) then $\beta_{sr} \approx \frac{\beta_{ml}}{\sqrt{1+0.3\sigma^2}} = \frac{2.29}{\sqrt{1+0.3*57.75}} = 0.534$ which is smaller than 2.29. Similarly, the regression coefficients (β_{ml}) of cycle 2= 0.79 obtained from multilevel modeling-scaled weights technique then $\beta_{sr} \approx \frac{\beta_{ml}}{\sqrt{1+0.3\sigma^2}} = \frac{0.79}{\sqrt{1+0.3*57.75}} = 0.18$ which is smaller than 0.79 and exactly similar to the estimated regression coefficient obtained from standard regression. This is one of the main reasons for this difference between the regression coefficients obtained from MM-SW and SR-RV estimation technique.

The MM-SW technique used multilevel pseudo maximum likelihood (MPML) to estimate the regression coefficients estimators where probability weights for each level unit were used, and the SR-RV technique used generalized estimating equations (GEE) based on quasi-likelihood function to estimate the regression coefficient estimators where overall probability weight were used [14, 25, 57]. Scaling of the probability weight was used in the MM-SW technique, whereas raw probability weight was used in the SR-RV estimation technique. Scaling of the weights had an influence on the estimation of the standard errors [25, 56, and 60]. The quasi-likelihood and pseudo likelihood approaches are different, but the pseudo likelihood approach may have complexity for non-normal data [40].

The findings based on our analyses of the longitudinal NPHS data indicate that the performance of the SR-RV estimation technique might be better than the MM-SW technique for analyzing longitudinal complex survey data.

Table 5.12 Estimates (Standard Errors) and their 95% confidence intervals based on the NPHS

	Multilevel modeling		Standard regression	
Covariates	Scaled weights		Robust(Bootstrap)	
	Estimate (SE)	95% C.I.	Estimate (SE)	95% C.I.
Intercepts	–21.89 (1.09)*	–24.03, –19.75	–5.30 (0.19)*	–5.69, –4.91
Age Groups				
18–44years (ref)				
45–64 years	2.29 (0.31)*	1.69, 2.89	0.64 (0.11)*	0.44, 0.85
65 years and above	5.00 (0.40)*	4.22–5.77	1.23 (0.14)*	0.96, 1.51
Location of Residence				
Urban (ref)				
Rural	0.37 (0.23)	–0.08, 0.83	–0.12 (0.07)	–0.26, 0.01
Sex				
Female (ref)				
Male	0.91 (0.49)	–0.05, 1.87	0.50 (0.16)*	0.12, 0.80
Education Levels				
University (ref)				
Secondary	1.09 (0.67)	–0.22, 2.39	0.38 (0.15)*	0.09, 0.67
Elementary	2.50 (0.57)*	1.42, 3.66	0.71 (0.18)*	0.36, 1.05
Household Income				
>$50,000 (ref)				
$30,000–$49,999	0.88 (0.40)*	0.10, 1.65	0.33 (0.11)*	0.11, 0.54
$15,000–$29,999	1.14 (0.40)*	0.36, 1.92	0.50 (0.14)*	0.23, 0.78
<$15,000	0.92 (0.41)*	0.12, 1.73	0.57 (0.16)*	0.26, 0.88
Physical Activity				
Yes (ref)				
No	0.14 (0.16)	–0.18, 0.46	–0.05 (0.05)	–0.15, 0.05
Body Mass Index				
BMI: <25 (ref)				
BMI: 25–29.9	0.63 (0.22)*	0.19, 1.07	0.06 (0.07)	–0.09, 0.20
BMI:>29.9	1.09 (0.29)*	0.52, 1.65	0.30 (0.11)*	0.08, 0.52

Cont'd Table 5.12

Covariates	Multilevel modeling Scaled weights		Standard regression Robust(Bootstrap)	
	Estimate (SE)	95% C.I.	Estimate (SE)	95% C.I.
Mother had diabetes				
No (ref)				
Yes	2.78 (0.58)*	1.64, 3.91	0.95 (0.14)*	0.68, 1.22
Father had diabetes				
No (ref)				
Yes	4.13 (0.84)*	2.48, 5.79	0.70 (0.16)*	0.38, 1.02
Time				
Cycle 1 (ref)				
Cycle 2	0.79 (0.19)*	0.42, 1.15	0.18 (0.05)*	0.08, 0.28
Cycle 3	1.56 (0.20)*	1.18, 1.96	0.32 (0.07)*	0.18, 0.46
Cycle 4	2.44 (0.26)*	1.93, 2.95	0.56 (0.07)*	0.42, 0.70
Cycle 5	3.50 (0.32)*	2.90, 4.14	0.75 (0.08)*	0.60, 0.91
Cycle 6	4.11 (0.34)*	3.44, 4.79	0.89 (0.08)*	0.73, 1.05
Interaction (Sex*household income)				
Male*$30,000–$49,999	–0.59 (0.49)	–1.55, 0.37	–0.29 (0.14)*	–0.56, –0.01
Male*$15,000–$29,999	–0.60 (0.61)	–1.80, 0.61	–0.32 (0.17)	–0.65, 0.01
Male*<$15,000	–0.46 (0.70)	–1.84, 0.91	–0.40 (0.20)*	–0.80, –0.002

* indicates p-value ≤ 0.05

5.3 Results based on Monte Carlo Simulation Technique

The third objective of this thesis was to investigate which statistical method was optimal for analyzing cross-sectional complex survey data using Monte Carlo simulation study. It is often challenging for applied researchers to find the appropriate statistical method to analyze the complex survey data. The performance of the MM-SW technique and the SR-RV estimation technique was assessed based on the empirical results obtained from the analyses of simulated cross-sectional complex survey data. To accomplish the assessment of performance between these two statistical techniques, a Monte Carlo simulation study was conducted to generate simulated data and analyze the simulated data using the MM-SW technique and the SR-RV estimation technique.

The sampling design for the Monte Carlo simulation technique was similar to Saskatchewan data. The simulated cross-sectional complex survey data were generated with the 100 and 1000 replications separately. In the Monte Carlo simulation technique, the RANTBL function in SAS® program was used to generate the simulated data for two independents variables: Body Mass Index (BMI) (<25 kg/m^2, ≥25kg/m^2) and EDUCATION (<secondary, ≥secondary) . Both independent variables were categorical (i.e. dichotomous). The logistic regression was used to generate the outcome variable of interest (type 2 diabetes (yes, no)) using above two independents variables: BMI (<25 kg/m^2, ≥25kg/m^2) and EDUCATION (<secondary, ≥secondary) . The detail procedures of generating simulated data using Monte Carlo simulation technique based on the Saskatchewan data was discussed in Section 4.3. Each of the simulated data sets with 100 and 1000 replications was analyzed using the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique where outcome variable was type 2 diabetes (yes, no) and the independent variables were BMI (<25 kg/m^2, ≥25kg/m^2) and EDUCATION (<secondary, ≥secondary). The logistic regression model with the given two independents variables was

Logit (Pr(diabètes = yes |x)) = $\hat{\beta}_0 + \hat{\beta}_{bmi} * BMI\left(\geq 25kg/m^2\right) + \hat{\beta}_{edu} * Education\left(< \text{sec}ondary\right)$.

The two statistical techniques (the MM-SW and the SR-RV estimation technique) were also used to analyze the observed Saskatchewan data (part of CHHS). The parameter estimates obtained from the analysis of observed Saskatchewan data was considered as a true parameter estimates. The assessment criteria to assess the performance of the MM-SW technique and the SR-RV estimation technique were bias of regression coefficients , percentage bias of regression coefficients, standardized bias of regression coefficients, means square errors (MSE), length of 95% confidence intervals, coverage of true regression coefficients in corresponding simulated 95% confidence intervals, and relative efficiency. The definition for each of assessment criteria was described in Table 3.3.1. The following Table 5.13 described the results of assessment criteria obtained from the Monte Carlo simulation study. It was mentioned in simulation procedure that the two groups of simulated data were generated: one with 100 replications and other one with 1000 replications. The results obtained from the analysis of the two groups of simulated data sets by applying multilevel modeling-scaled weights technique and standard regression-robust variance estimation technique were discussed below.

Results based on simulated data with 1000 replications

Results based on the analysis of simulated data with the 1000 replications using the multilevel modeling–scaled weights technique and the standard regression–robust variance estimation technique indicated that the biases, percentage biases, standardized biases for the regression coefficients of BMI (<25 kg/m^2, ≥ 25kg/m^2) were higher in the multilevel modeling-scaled weights compared to the standard regression-robust variance estimation technique. The biases, percentage biases, standardized biases for the regression coefficients of education level (<secondary, ≥secondary) were almost similar between the multilevel modeling-scaled weights technique and the standard regression-robust variance estimation technique (Table 5.13). Means

square errors, length of 95% C.I. of regression coefficients were also similar between the MM-SW technique and the SR-RV estimation technique (Table 5.13).

The coverage of the true regression coefficients for both independent variables in the corresponding simulated 95% confidence intervals was higher in the MM-SW technique compared to the SR-RV estimation technique. The efficiency of the MM-SW technique and the SR-RV technique was similar according to the calculation of relative efficiency. The results from the bias (biases, percentage biases, standardized biases) of regression coefficients indicated that the performance of the SR-RV estimation technique was better than the MM-SW technique based on simulated data with 1000 replications.

The results from other criteria such as MSE, length of 95% C.I. of regression coefficients and relative efficiency indicated that the performance of the MM-SW technique and the SR-RV estimation technique was similar. The performance of the MM-SW technique was slightly better compared to the SR-RV estimation technique based on coverage of true regression coefficients in corresponding simulated 95% confidence interval to analyze complex survey data.

Results based on simulated data with 100 replications

Based on the analysis of simulated data with the 100 replications using the MM-SW technique and the SR-RV estimation technique, results indicated that the biases , percentage biases, standardized biases of regression coefficients for both independents variables : (BMI (<25 kg/m^2, ≥25kg/m^2) and EDUCATION (<secondary, ≥secondary) were higher in the MM-SW technique compared to the SR-RV estimation technique . Means square errors, length of 95% C.I. of regression coefficients were lower in the MM-SW technique compared to the SR-RV estimation technique (Table 5.13).

The coverage of the true regression coefficients for the both independent variables in the corresponding 95% confidence intervals obtained from the simulated data were higher in the MM-SW technique compared to the SR-RV estimation technique (Table 5.13). According to the calculation of relative efficiency, the efficiency of MM-SW technique was higher than SR-RV technique (Table 5.13). The results based on bias of regression coefficients from the analysis of simulated data with 100 replications indicated that the performance of the SR-RV estimation technique was better compared to the MM-SW technique. The obtained results based on MSE, length of 95% C.I. of regression coefficients and coverage of true regression coefficients in corresponding simulated 95% C.I. also indicated that the performance of the MM-SW technique was better than the SR-RV estimation technique.

Results based on simulated data with 100 and 1000 replications using multilevel modeling-scaled weights technique

Based on the empirical results obtained from the analysis of the simulated data with 100 and 1000 numbers of replications using the MM-SW technique, results pointed out that the standard errors of regression coefficients were smaller when the numbers of replications were increased from 100 to 1000. The biases, percentage biases, standardized biases, means square errors for the regression coefficients of both covariates (BMI (<25 kg/m^2, ≥ 25kg/m^2) and EDUCATION ($<$secondary, $\geq$secondary) were lower in the simulated data with 1000 replications compared to the simulated data with 100 replications using multilevel MM-SW technique. This result indicated that the higher number of simulations reduced the bias for the regression coefficients. The MSE were also lower in simulated data with 1000 replications compared to simulated data 100 replications data. The coverage of the true regression coefficients obtained from the observed Saskatchewan data for both independent variables in the corresponding simulated 95% confidence intervals obtained from simulated data with 1000 replications was higher compared to the simulated data with 100

replications in multilevel modeling-scaled weights technique. Lengths of 95% C.I. of regression coefficients obtained from the analysis of both simulated data with the 100 and 1000 replications were similar in MM-SW technique. The results obtained from the analysis of simulated complex survey data sets using MM-SW indicated that the data with higher numbers of replications provides more reliable and consistent parameters estimates.

Results based on simulated data with 100 and 1000 replications using standard regression-robust variance estimation technique

The empirical results obtained from the analysis of the simulated data with 100 and 1000 replications using the SR-RV estimation technique indicated that the standard errors of regression coefficients were smaller when the numbers of replications were increased from 100 to 1000. The biases, percentage biases, standardized biases of regression coefficients for the body mass index (<25 kg/m^2, ≥ 25kg/m^2) were lower in the simulated data with 1000 replications compared to the simulated data with 100 replications using SR-RV estimation technique. The biases, percentage biases and standardized biases for the regression coefficients of the education (<secondary, ≥secondary) were higher in the simulated data with 1000 replications compared to the simulated data with 100 replications using SR-RV estimation technique. The MSE were lower in simulated data with 1000 replications compared to simulated data with 100 replications. The coverage of the true regression coefficients obtained from the observed Saskatchewan data for both independent variables in the corresponding simulated 95% confidence intervals obtained from simulated data with 1000 replications was higher compared to the simulated data with 100 replications in SR-RV estimation technique. Lengths of 95% C.I. of regression coefficients for both covariates obtained from the analysis of both simulated data with the 100 and 1000 replications were similar in SR-RV estimation technique.

Based on the comparison of results of assessment criteria from the analysis of simulated data with two types of number of replications or sample sizes such as 100 and 1000 using both statistical techniques, the results from the analysis of data with the higher replications (1000) provided the precise results compared to the results obtained from the analysis of data with the lower replications (100).

To summarize based on 1000 replications, the five assessment criteria: bias (variation of bias of regression coefficients, percentage bias of regression coefficients and standardize bias of regression coefficients), means square errors (MSE), coverage of the true regression coefficients in simulated 95% C.I. , length of 95% C.I. of regression coefficients, and relative efficiency to assess which method is appropriate to analyze the cross-sectional complex survey data. The three assessment criteria such MSE, coverage of the true regression coefficients in simulated 95% C.I. and length of 95% C.I. of regression coefficients did not reveal that the two analytical techniques under investigative would provide different results. However, we did observe that based on bias, SR-RV estimation technique is an appropriate method compared to MM-SW technique.

Table 5.13 Results for assessment criteria to compare the performance of the MM-SW technique and the SR-RV estimation technique based on Monte Carlo simulation

Evaluation criteria	variables	Multilevel modeling-scaled weights		Standard regression-robust variance	
		1000 simulated data sets	100 simulated data sets	1000 simulated data sets	100 simulated data sets
Bias of regression coefficients	BMI: <25 kg/m^2(ref) ≥25kg/m^2	0.00820	0.0925	0.0009	0.0829
	EDUCATION: ≥secondary (ref) <secondary	0.02560	0.0491	0.0330	0.0059
Percentage bias of regression coefficients	BMI: <25 kg/m^2(ref) ≥25kg/m^2	1.08%	12.22%	0.11%	10.85%
	EDUCATION: ≥secondary (ref) <secondary	5.50%	10.54%	8.09%	0.55%
Standardize bias of regression coefficients	BMI: <25 kg/m^2(ref) ≥25kg/m^2	5.46%	40.08%	0.587%	34.79%
	EDUCATION: ≥secondary (ref) <secondary	17.21%	30.49%	22.24%	3.21%

Cont'd Table 5.13

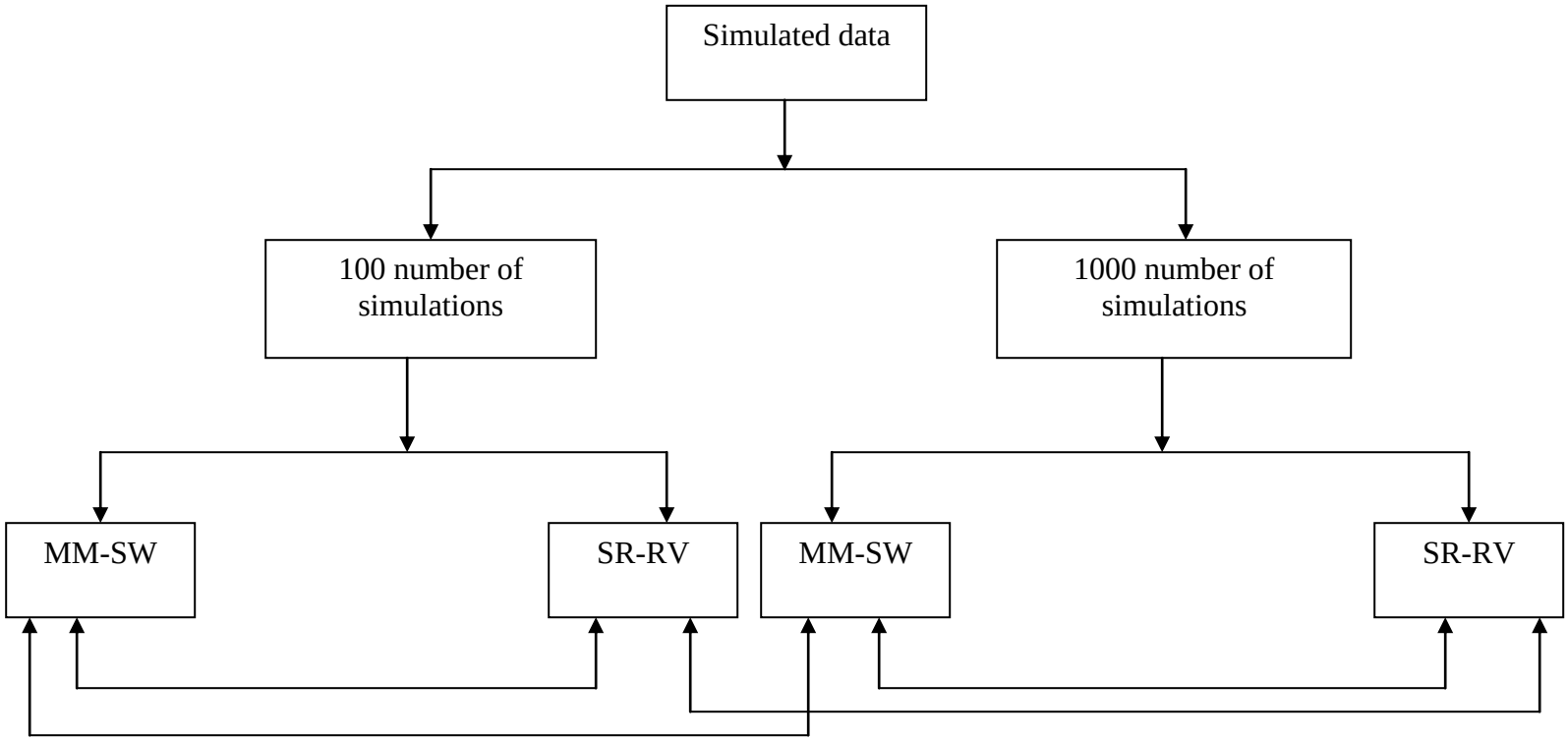
Evaluation criteria	variables	Multilevel modeling-scaled weights		Standard regression-robust variance	
		1000 simulated data sets	100 simulated data sets	1000 simulated data sets	100 simulated data sets
Means square errors	BMI:				
	<25 kg/m^2(ref)				
	≥25 kg/m^2	0.0225	0.0618	0.0212	0.0632
	EDUCATION:				
	≥secondary (ref)				
	<secondary	0.0228	0.0284	0.0232	0.0296
Coverage of the true regression coefficients in simulated 95% C.I.	BMI:				
	<25 kg/m^2(ref)				
	≥25 kg/m^2	91%	78%	83%	72%
	EDUCATION:				
	≥secondary (ref)				
	<secondary	89%	78%	81%	72%
Average length of 95% confidence intervals of regression coefficients	BMI:				
	<25 kg/m^2(ref)				
	≥25.kg/m^2	0.6037	0.5859	0.5938	0.5898
	EDUCATION:				
	≥secondary (ref)				
	<secondary	0.5616	0.5494	0.5766	0.5775

Cont'd Table 5.13

Evaluation criteria	variables	Multilevel modeling-scaled weights		Standard regression-robust variance	
		1000 simulated data sets	100 simulated data sets	1000 simulated data sets	100 simulated data sets
Relative efficiency	BMI:				
	<25 kg/m^2(ref)				
	≥25 kg/m^2	1.06	0.94	1.00	1.00
	EDUCATION:				
	≥secondary (ref)				
	<secondary	1.00	0.88	1.00	1.00

The statistical formula for each criteria are shown in Table 3.3.1

Figure 5.3 Comparison between MM-SW technique and SR-RV technique based on the results obtained from the analysis of simulated data with two sample sizes



Goodness of fit test for the logistic regression model based on Monte Carlo simulation study

A number of 1000 simulated data sets, each of sample size 1,731 were generated. The binary outcome variable was type 2 diabetes and explanatory variables were education and BMI. Both explanatory variables were categorical. The logistic regression model with two explanatory variables was fitted for each of 1000 simulated data sets in this Monte Carlo simulation study. In order to test the goodness of fit for the model, 1000 simulated data sets were divided into four batches 250 each batch. Random sample of the twenty data sets were selected from each batch to create four groups. The *estat gof* STATA code was used to estimate the goodness of fit statistic and the corresponding p-value. Based on the results of goodness of fit test for each data set in each group, only one data set was not fitted well in the first group and another data set was not fitted well in third group (Table 5.14.1, Table 5.14.2, Table 5.14.3, Table 5.14.4). This results indicated that almost all data sets in each group were fitted the logistic regression model adequately. Hence, based on the results of goodness of fit test it can be concluded that the logistic regression models fitted the simulated data adequately (Table 5.14.1, Table 5.14.2, Table 5.14.3, Table 5.14.4).

Table 5.14.1 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 1)

Data sets	$\chi^2(9)$	P-value
1	1.38	0.9979
2	1.56	0.9966
3	3.50	0.9411
4	3.31	0.9505
5	0.61	0.9999
6	0.99	0.9995
7	4.17	0.9001
8	0.02	1.00
9	**45.44**	**0.00**
10	5.05	0.8297
11	0.60	0.9999
12	2.87	0.9692
13	2.86	0.9695
14	0.29	1.00
15	2.91	0.9676
16	0.08	1.00
17	3.12	0.9592
18	1.53	0.9969
19	1.40	0.9978
20	5.26	0.8115

Table 5.14.2 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 2)

Data sets	$\chi^2(9)$	P-value
1	0.38	1.00
2	0.87	0.9979
3	0.01	1.00
4	0.10	1.00
5	8.17	0.5172
6	0.41	1.00
7	0.41	1.00
8	0.41	1.00
9	0.08	1.00
10	0.27	1.00
11	0.00	1.00
12	0.08	1.00
13	0.02	1.00
14	2.43	0.9828
15	0.41	1.00
16	0.61	0.9999
17	2.83	0.9707
18	0.01	1.00
19	0.27	1.00
20	0.05	1.00

Table 5.14.3 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 3)

Data sets	$\chi^2(9)$	P-value
1	1.78	0.9945
2	0.43	1.00
3	0.87	0.9997
4	10.18	0.3365
5	0.28	1.00
6	0.59	0.9999
7	0.34	1.00
8	**22.20**	**0.008**
9	0.04	1.00
10	0.75	0.9998
11	5.83	0.7567
12	0.51	1.00
13	0.03	1.00
14	0.00	1.00
15	0.86	0.9997
16	9.59	0.3848
17	0.05	1.00
18	2.51	0.9805
19	0.53	1.00
20	0.04	1.00

Table 5.14.4 Goodness of fit statistic with p-value based on 20 simulated data sets (Group 4)

Data sets	$\chi^2(9)$	P-value
1	0.23	1.00
2	11.32	0.2542
3	0.32	1.00
4	3.09	0.9604
5	1.23	0.9987
6	0.44	1.00
7	4.87	0.8453
8	0.09	1.00
9	6.31	0.7087
10	1.11	0.9991
11	1.53	0.9969
12	0.55	1.00
13	0.01	1.00
14	0.51	1.00
15	0.45	1.00
16	0.15	1.00
17	0.12	1.00
18	3.43	0.9445
19	2.76	0.9731
20	0.07	1.00

5.4 Interpretation of results

For cross-sectional complex survey data, the findings from the Monte Carlo simulation technique indicated that the MM-SW technique and SR-RV estimation technique performed equally well to analyze cross-sectional complex survey data. However, the interpretation of the results is based on the SR-RV estimation technique, as other researchers have suggested that it is appropriate to use bootstrap variance estimates because this technique accounts for design effects (stratification and clustering) more accurately as compared to MM-SW technique which still has some deficiencies because weights are not available at each level for publicly used complex survey data sets. Statistics Canada develops bootstrap weights for surveys based on multi-stage complex design for the purposes of formulating correct inferences [106]. Hence, the estimated regression coefficients from standard regression – robust variance technique and their robust standard errors were used for interpretation purposes. These results are presented in section 5.4.1.

For longitudinal complex survey data, the findings from the analyses of the NPHS data were reliable compared to the results based on MM-SW technique. Because the former technique accounted for unequal probability of selection, design effects and it also accounted for within-subject correlation. The MM-SW technique did not have provision to account for within-subject correlation when analyzing longitudinal complex survey data. The results obtained from the analyses of the longitudbnal complex survey data (NPHS) using the SR-RV estimation technique were used for interpretation. Hence, the estimated regression coefficients and their robust standard errors were used for interpretation purposes. The interpretation of the results obtained from the SR-RV estimation technique based on the longitudinal complex survey (NPHS) was presented in section 5.4.2.

The interpretations of estimated regression coefficients and odds ratios are different between MM-SW and SR-RV estimation technique because MM-SW is a subject-specific method and SR-RV is a population averaged model. For example, interpretation of OR =2.5 for obese individual compared to normal individual, obtained from MM-SW: The odds of developing type 2 diabetes for an obese individual is 2.5 times higher compared to a normal individual controlling for the other covariates. Interpretation of OR =1.8 for obese people compared to normal people, obtained from SR-RV: The odds of developing type 2 diabetes among obese people is 1.8 times higher compared to the normal people.

5.4.1 Interpretation of results based on standard regression-robust variance estimation technique from the cross-sectional complex survey: CHHS

Participants who were unemployed (OR: 1.82, 95% C.I. 1.01, 3.25) were more likely to have self-reported, physician-diagnosed type 2 diabetes compared with the participants who had full-time jobs controlling for the other covariates. Participants who were homemakers (OR: 1.99, 95% C.I. 1.16, 3.42) or retired (OR: 1.80, 95% C.I. 0.89, 3.67) were also more likely to have self-reported, physician-diagnosed type 2 diabetes compared with the participants who had full-time jobs controlling for the other covariates. The odds of developing self-reported, physician-diagnosed type 2 diabetes were 16% lower among the participants who worked part-time or were students (OR: 0.84, 95% C.I. 0.57, 1.30) compared with the participants who had full-time jobs. The probability of developing self-reported, physician-diagnosed type 2 diabetes was higher among obese (BMI>29.9 kg/m^2) participants (OR: 3.46, 95% C.I. 2.53, 4.76) than among participants with a normal weight (BMI<25 kg/m^2). Similarly, the probability of developing self-reported, physician-diagnosed type 2 diabetes was higher among overweight (BMI = 25 – 29.9

kg/m^2) participants (OR: 1.82, 95% C.I. 1.16, 2.86) than among participants with a normal weight (BMI<25 kg/m^2).

In the final model based on CHHS, there was a significant interaction between sex and age group associated with the development of self-reported, physician-diagnosed type 2 diabetes. Odd ratios were calculated for the interaction term using Hosmer and Lemeshow's approach [27]. The estimated odd ratios and their 95% confidence interval were shown in Table 5.16. The results indicated that the risk of type 2 diabetes were significantly increased among male and female with both (45-64 years and 65-74 years) age groups with exception of 45-64 years female group. When comparing between male and female, male participants had higher risk of type 2 diabetes compared to female participants for both (45-64 years and 65-74 years) age groups.

The prediction of the probability of type 2 diabetes and the risk factors can be summarized using the following final logistic regression models based on SR-RV estimation technique with the main effects and the interaction terms:

Logit [Pr((Type 2 Diabetes)$_i$ =1)] = –4.89 + 0.25*(45 – 64 years)$_i$ + 0.74*(65 – 74 years)$_i$ + 0.20*(Urban)$_i$ – 0.41*(male)$_i$ + 0.40*(secondary)$_i$ + 0.69*(elementary)$_i$ + 0.17*($25,000–$49,999)$_i$ + 0.17*($12,000 – $24,999)$_i$ - 0.04*(<$12,000)$_i$ – 0.17*(part-time/students)$_i$ +0.60*(unemployment)$_i$ + 0.69*(homemaker)$_i$ + 0.59*(retired)$_i$ +0.36*(no physical activity)$_i$ + 0.60*(bmi:25 – 29.9 kg/m^2)$_i$ + 1.24*(bmi>29.9 kg/m^2) + 1.21*(45–64 years-male)$_i$ + 0.84* (65 – 74 years – male)$_i$.

Table 5.15 Odds ratios (OR) and 95% confidence intervals (95% C.I.) based on the SR-RV estimation technique using the CHHS

Covariates	Estimates (SE)	Odds Ratio (95% C.I.)
Age		
18–44years (ref)		
45–64 years	0.25 (0.21)	1.28 (0.85 – 1.93)
65 years and above	0.74(0.33)*	2.10 (1.11 – 3.94)
Location of Residence		
Rural (ref)		
Urban	0.20 (0.15)	1.22 (0.92 – 1.63)
Sex		
Female (ref)		
Male	–0.41 (0.34)	0.66 (0.34 – 1.30)
Education Levels		
University (ref)		
Secondary	0.40 (0.26)	1.49 (0.90 – 2.048)
Elementary	0.69 (0.34)*	1.99 (1.02 – 3.94)
Household Income		
>$50,000 (ref)		
$25,000–$49,999	0.17 (0.32)	1.19 (0.63 – 2.23)
$12,000–$24,999	0.17 (0.23)	1.19 (0.76 – 1.88)
<$12,000	-0.04 (0.24)	0.96 (0.61 – 1.54)
Employment Status		
Full-time (ref)		
Part-time/students	–0.17 (0.24)	0.84 (0.57 – 1.30)
Unemployment	0.60 (0.30)*	1.82 (1.01 – 3.25)
Homemaker	0.69 (0.27)*	1.99 (1.16 – 3.42)
Retired	0.59 (0.36)	1.80 (0.89 – 3.67)
Physical Activity		
Yes (ref)		
No	0.36 (0.20)	1.43 (0.97 – 2.12)
Body Mass Index		
BMI: <25 (ref)		
BMI: 25.0–29.9	0.60 (0.23)*	1.82 (1.16 – 2.86)
BMI:>29.9	1.24 (0.16)*	3.46 (2.53 – 4.76)
Interaction		
Age groups * sex		
18–44yrs*female (ref)		
45–64yrs*male	1.21 (0.38)*	3.35 (1.60 – 7.03)
65–74yrs*male	0.84 (0.35)*	2.32 (1.17 – 4.57)

* indicates P-value ≤0.05

Table 5.16 Calculation of odd ratios for interaction terms based on CHHS

Effect	Among	OR	95% C.I.
Age group			
45-64 years	male	4.31	1.49 – 12.47
65-74 years	male	4.85	1.43 – 16.44
45-64 years	female	1.28	0.85 – 1.94
65-74 years	female	2.10	1.10 – 4.00

5.4.2 Interpretation of the results obtained from SR-RV estimation technique based on the longitudinal complex survey: NPHS

Participants in the age group 45–64 years were more likely (OR: 1.90, 95% C.I. 1.55–2.34) to develop self-reported, physician-diagnosed type 2 diabetes compared with the younger participants in the age group 18–44 years after controlling for the other covariates. In the same way, participants in the age group 65 years and over were more likely (OR: 3.43, 95% C.I. 2.60–4.51) to develop self-reported, physician-diagnosed type 2 diabetes than were younger participants in the age group 18–44 years after controlling for the other covariates.

Participants with elementary school or less education were more likely (OR: 2.03, 95% C.I. 1.44, 2.86) to develop self-reported, physician-diagnosed type 2 diabetes compared with those who had university degrees. Similarly, participants with secondary education (OR: 1.47, 95% C.I. 1.1, 1.96) were more likely to develop self-reported, physician-diagnosed type 2 diabetes compared with those who had university degrees after controlling for the other covariates.

The risk of developing self-reported, physician-diagnosed type 2 diabetes was higher (OR: 1.35, 95% C.I. 1.09, 1.69) among obese (BMI>29.9 kg/m^2) participants compared with the participants with a normal weight (BMI<25 kg/m^2) after controlling for the other covariates. The risk of developing self-reported, physician-diagnosed type 2 diabetes was also higher (OR: 1.06, 95% C.I. 0.91, 1.23) among overweight (BMI: 25–29.9 kg/m^2) participants than among participants with a normal weight (BMI<25 kg/m^2) after controlling for the other covariates.

Participants who reported that their birth mother had type 2 diabetes were more likely (OR: 2.59, 95% C.I. 1.98, 3.40) to develop self-reported, physician-diagnosed type 2 diabetes compared with participants who did not report that their birth mother had type 2 diabetes. Participants who reported that their birth father had type 2 diabetes were more likely (OR: 2.01, 95% C.I. 1.46, 2.77) to develop self-reported, physician-diagnosed type 2 diabetes compared with participants who did not report that their birth father had type 2 diabetes.

There was a significant interaction between sex and household income associated with the development of self-reported, physician-diagnosed type 2 diabetes. Odd ratios were calculated for the interaction term based on Hosmer and Lemeshow's approach [27]. Estimated odd ratios (OR) were shown in Table 5.18. Odd ratios for interaction terms indicated that lower household income significantly increased the risk of type 2 diabetes among female participants but not true for male participants.

The prediction of the probability of type 2 diabetes and the risk factors can be summarized using the following final logistic regression models from SR-RV estimation technique that are based on the main effects and interaction terms:

$$\text{Logit}\,[\Pr((\text{Type 2 Diabetes})_{ij}=1)] = -5.30 + 0.64*(45-64\ \text{years})_{ij} + 1.23*(65\ \text{and above years})_{ij} - 0.12*(\text{Rural})_{ij} + 0.50*(\text{male})_{ij} + 0.38*(\text{secondary})_{ij} + 0.71*(\text{elementary})_{ij} + 0.57*(<\$15{,}000)_{ij} + 0.50*(\$15{,}000-\$29{,}999)_{ij} + 0.33*(\$30{,}000-\$49{,}999)_{ij} - 0.05*(\text{physical activity})_{ij} + 0.06*(\text{bmi:}25-29.9\ \text{kg/m}^2)_{ij} + 0.30*(\text{bmi}>29.9\ \text{kg/m}^2) + 0.95*(\text{Birth mother had diabetes})_{ij} + 0.70*(\text{Birth father had diabetes})_{ij} + 0.18*(\text{Cycle 2})_{ij} + 0.32*(\text{Cycle 3})_{ij} + 0.56*(\text{Cycle 4})_{ij} + 0.75*(\text{Cycle 5})_{ij} + 0.89*(\text{Cycle 6})_{ij} - 0.29*(\$30{,}000-\$49{,}999-\text{male})_{ij} - 0.32*(\$15{,}000-\$29{,}999-\text{male})_{ij} - 0.40*(<\$15{,}000-\text{male})_{ij}.$$

Table 5.17 Odd ratios (95% confidence intervals) based on the SR-RV estimation technique using the NPHS

Covariates	Estimates (SE)	Odd Ratio (95% C.I.)
Age Groups		
18–44years (ref)		
45– 64 years	0.64 (0.11)*	1.90 (1.55 – 2.34)
65 years and above	1.23 (0.14)*	3.43 (2.60 – 4.51)
Location of Residence		
Urban (ref)		
Rural	–0.12 (0.07)	0.88 (0.77 – 1.01)
Sex		
Female (ref)		
Male	0.50 (0.16)*	1.64 (1.21 – 2.23)
Education Levels		
University (ref)		
Secondary	0.38 (0.15)*	1.47 (1.1 – 1.96)
Elementary	0.71 (0.18)*	2.03 (1.44 – 2.86)
Household Income		
>$50,000 (ref)		
$30,000–$49,999	0.33 (0.11)*	1.39 (1.12 – 1.72)
$15,000–$29,999	0.50 (0.14)*	1.66 (1.26 – 2.17)
<$15,000	0.57 (0.16)*	1.77 (1.30 – 2.42)
Physical Activity		
Yes (ref)		
No	–0.05 (0.05)	0.95 (0.86 – 1.05)
Body Mass Index		
BMI: <25 (ref)		
BMI: 25–29.9	0.06 (0.07)	1.06 (0.91 – 1.23)
BMI:>29.9	0.30 (0.11)*	1.35 (1.09 – 1.69)
Mother had diabetes		
No (ref)		
Yes	0.95 (0.14)*	2.59 (1.98 – 3.40)
Father had diabetes		
No (ref)		
Yes	0.70 (0.16)*	2.01 (1.46 – 2.77)
Time		
Cycle 1 (ref)		
Cycle 2	0.18 (0.05)*	1.20 (1.10 – 1.32)
Cycle 3	0.32 (0.07)*	1.38 (1.20 – 1.59)
Cycle 4	0.56 (0.07)*	1.75 (1.52 – 2.01)
Cycle 5	0.75 (0.08)*	2.13 (1.82 – 2.49)
Cycle 6	0.89 (0.08)*	2.43 (2.10 – 2.84)

Cont'd Table 5.17

Covariates	Estimates (SE)	Odds Ratio (95% C.I.)
Interaction (Sex*household income)		
Male*$30,000–$49,999	–0.29 (0.14)*	0.67 (0.45 – 1.00)
Male*$15,000–$29,999	–0.32 (0.17)	0.73 (0.52 – 1.01)
Male*<$15,000	–0.40 (0.20)*	0.75 (0.57 – 0.99)

* indicates P-value ≤0.05

Table 5.18 Calculation of odd ratios for interaction terms based on NPHS

Effect	Among	OR	95% C.I.
Household income			
<$15,000/year	male	1.20	0.63 – 2.22
$15,000 - $29,999/year	male	1.20	0.69 – 2.07
$30,000-$49,999/year	male	1.04	0.67 – 1.62
<$15,000/year	female	1.77	1.29 – 2.42
$15,000 - $29,999/year	female	1.65	1.25 – 2.17
$30,000-$49,999/year	female	1.39	1.12 – 1.73

CHAPTER 6

DISCUSSION AND CRITIQUE OF RESULTS

6.1 Introduction

Statistical methods are well established for data from a simple random sampling (SRS) framework in which the observations are independent of each other. Observations may not be independent of each other in real-life, complex surveys that are based on multistage design. Traditional statistical methods, which assume observations are independent of each other, are not appropriate for analysis of such surveys. The MM-SW technique and the SR-RV estimation technique are commonly used statistical techniques to analyze data obtained from cross-sectional and longitudinal complex surveys. In this thesis, these two statistical methods were compared by using them to analyze binary data obtained from cross-sectional (CHHS) and longitudinal (NPHS) complex surveys. The outcome variable of interest was type 2 diabetes. A Monte Carle simulation study was also conducted to assess and identify the more suitable statistical method between these two methods for cross-sectional complex survey data.

In section 6.2, a comparison between these two statistical techniques was made based on the results of the analyses of the data obtained from cross-sectional complex surveys. In section 6.3, a comparison between these two statistical techniques was made based on the results of the analyses of the data obtained from longitudinal complex surveys. Finally, a comparison between these two statistical techniques was made based on the results obtained from the Monte Carlo simulation study in section 6.4.

6.2 Objective 1: To compare the multilevel modeling–scaled weights(MM-SW) technique and the standard regression–robust variance (SR-RV) estimation technique by analyzing cross-sectional complex survey data.

The first objective of this thesis was to compare these two statistical techniques based on the estimated standard errors and 95% confidence intervals for the estimated regression coefficients obtained from the analysis of the CHHS. The results from the analysis of the CHHS data indicated that the estimated regression coefficients were different between the two techniques. With the exception of a few variables, the regression coefficient estimates obtained from the MM-SW technique were higher compared with the SR-RV estimation technique (Table 5.6). This was expected, as is explained below. In the MM-SW technique, the regression coefficients were calculated using multilevel pseudo maximum likelihood (MPML) with numerical integration via adaptive quadrature, where the probability weight for each level were incorporated in the MPML [58]. It is difficult to have a closed-form of the marginal likelihood in generalized linear mixed models or multilevel models. Marginal likelihood is a joint probability of all observed responses, given the covariates. Gauss–Hermite or ordinary quadrature is often used to evaluate and maximize the marginal likelihood for parameter estimation. Adaptive quadrature is an approximate method, and it can be used to approximate the marginal likelihood. Adaptive quadrature is more efficient at approximating the marginal likelihood than ordinary quadrature or Gauss–Hermite quadrature is at estimating the parameters, and it can be implemented in GLLAMM in STATA software program [87, 88]. Non-responses and unequal probabilities of selection occur at each level in multistage complex surveys. The effects of the non-response and unequal probability of selection are taken into account by probability weight variables at the analysis stage. In the multilevel modeling technique, separate probability weight for sampling units at each level of data are required to take into account these

design effects of complex surveys. For example, let p_k be the probability of selection for level 2 units (PSU level in the CHHS) and ω_k (where $\omega_k = \frac{1}{p_k}$) be the corresponding probability weights for level 2 units. Let $p_{i|k}$ be the probability of selection for a level 1 unit (individual in the CHHS) within level 2 (PSU in the CHHS) and $\omega_{i|k}$ (where $\omega_{i|k} = \frac{1}{p_{i|k}}$) be the corresponding probability weights for level 1 units. These probability weights for each level were incorporated in the multilevel pseudo maximum likelihood (MPML) approach to determine the parameter estimates.

On the other hand, the overall single level probability weight were incorporated in pseudo maximum likelihood (PML) to take into account the non-responses and unequal probability of selection for sampling units in the SR-RV estimation technique. This might be the main reason for the differences in the regression coefficients and their standard errors between the two techniques. In the analysis of data obtained from cross-sectional complex surveys, the multilevel modeling technique assumes that observations are dependent on each other, but standard logistic regression assumes that observations are independent of each other [49].

The relationship ($\beta_{sr} \approx \frac{\beta_{ml}}{\sqrt{1+0.3\sigma^2}}$) between the regression coefficients obtained from multilevel models (random-effects models) and standard regression indicated that regression coefficient estimates can be higher in the MM-SW technique than in the SR-RV estimation technique. This is one of the main reasons of difference in the values of regression coefficients between two techniques. In the MM-SW technique, the probability weight variable was used for each level, whereas a single probability weight variable was used in SR-RV estimation technique. If the weight variable for any level of multistage complex survey data is not available,

then the multilevel modeling technique assumes that this weight variable is equal to one for that level of complex survey data. This might be one of the reasons for the differences in the regression coefficient estimators between the two techniques.

The estimated standard errors of the estimated regression coefficients using the MM-SW technique were smaller compared with the SR-RV estimation technique, based on the analysis of the CHHS. The bootstrap variance estimation technique was used to estimate the standard errors of the parameter estimates in the SR-RV estimation technique, and sandwich variance estimators were used to estimate the standard errors of the parameter estimates in the MM-SW technique. The subpopulation might be the reason for this difference in the standard errors. The impact of clustering was taken into account by including additive random effects in the multilevel modeling, which can produce significantly underestimated standard errors [42]. Scaling of the level 1 weight has an influence on the parameter and their standard error estimations [57, 60]. Scaling of the probability weight might be another reason for the differences in the parameter estimates between the two statistical techniques.

Based on the results obtained from the analysis of the CHHS, it is difficult to recommend one technique as preferable for analyzing complex survey data. A Monte Carlo simulation study was conducted to determine the preferable statistical method. The preferable statistical method for analyzing cross-sectional complex survey data can be determined based on the results of the Monte Carlo simulation technique.

6.2.1 Prevalence of self-reported, physician-diagnosed type 2 diabetes and its risk factors among Canadians

Type 2 diabetes is a major health burden for Canadians, and it is also a well-known cause of heart disease, blindness and kidney failure. In this thesis, the prevalence of self-reported, physician-diagnosed type 2 diabetes and its risk factors among Canadian adults were studied, based on Canadian heart health surveys (CHHS). The prevalence of type 2 diabetes was higher in New Brunswick (5.5%, 95% C.I. 4.7–6.4), Newfoundland and Labrador (5.4%, 95% C.I. 4.8–6.2), and Saskatchewan (5.4%, 95% C.I. 3.6–8.0) compared with other provinces (Table 5.4). Based on the analyses of the CHHS using the multilevel modeling–scaled weights technique, the statistically significant predictors associated with the development of self-reported, physician-diagnosed type 2 diabetes (yes, no) were household income (<$12,000; $12,000–$24,999; $25,000–$49,999; >$50,000), employment status (full-time, part-time/student, unemployed, homemaker, retired), physical activity (yes, no) and body mass index (BMI<25 kg/m^2, BMI = 25–29.9kg/m^2, BMI>29.9kg/m^2). An interaction between age (18–44 years, 45–64 years, 65 years and above) and sex (male, female) was significantly associated with the development of self-reported, physician-diagnosed type 2 diabetes.

Several studies, including Canadian studies, have been conducted to determine the relationship between household income and the prevalence of type 2 diabetes. The study findings indicated that people with lower incomes are significantly associated with type 2 diabetes, which is similar to this study finding [89-91]. The probable reasons might be that people with lower incomes consume less nutritive food and have less access to fitness clubs because they cannot afford the membership cost.

Few studies have been conducted to establish the relationship between employment or occupational status and the prevalence of type 2 diabetes [92]. This study found that employment status made a significant contribution to whether Canadian participants developed type 2 diabetes, a result that supported the findings of previous studies. Several studies have indicated that obesity was a significant predictor of type 2 diabetes among Canadian women and the Métis of western Canada [17, 93]. A study in Canada found that obese (BMI>29.9) women were more likely to have type 2 diabetes [17]. The relationship between BMI and the prevalence of type 2 diabetes has been studied extensively. Being overweight (BMI: 25–29.9 kg/m^2) or obese (BMI>29.9kg/m^2) is significantly associated with the prevalence of type 2 diabetes [72, 94, 95]. This study findings were consistent with other findings. A Canadian study reported that physical activity was a significant predictor of type 2 diabetes, which this study findings supported [96]. Some studies found a weak relationship between physical activity and the prevalence of type 2 diabetes after adjusting the other covariates, such as BMI and gender [97]. Age and sex are significant risk factors among Iranian and Canadian adults [74, 93]. Canadian women over 40 and with a low SES have a higher prevalence of type 2 diabetes, but Canadian men do not [90]. These study findings found a combined effect of age and sex on the prevalence of type 2 diabetes among Canadian residents.

6.3 Objective 2: To compare the MM-SW technique and the SR-RV estimation technique for analyzing longitudinal complex survey data

The second objective of this thesis was to compare the performance of the SR-RV estimation technique (GEE-Liang and Zeger with bootstrap variance) and the MM-SW technique for longitudinal complex survey data. Based on the analyses of the NPHS data, the estimated regression coefficients obtained from the MM-SW technique were larger than the

estimated regression coefficients obtained from the standard regression (GEE)–robust variance estimation technique which is expected. The relationship between the regression coefficients obtained from MM-SW technique and SR-RV estimation technique indicated that the regression coefficients obtained from MM-SW technique were higher compared with standard regression-robust variance estimation technique. The estimated regression coefficients obtained from MM-SW were larger than the regression coefficients obtained from SR-RV estimation technique which is the agreement of the relationship between regression coefficients obtained from the MM-SW technique and SR-RV estimation technique. The standard errors of the estimated regression coefficients obtained from the MM-SW technique were also larger than the standard errors of the estimated regression coefficients obtained from the SR-RV technique .The 95% confidence intervals of the estimated regression coefficients were also wider for the MM-SW technique compare with the SR-RV estimation technique.

Multilevel modeling is known as subject-specific (SS) or random-effects modeling, and standard regression (GEE) is known as population-averaged or marginal modeling [38]. Several studies have indicated that the regression coefficient estimates were larger for the MM-SW technique compared with the regression coefficient estimates from GEE [10, 11, 21, 26]. This study results support those findings. The MM-SW technique has more computational problems, such as convergence issues, compared with the SR-RV estimation technique [20]. As a result, the SR-RV estimation technique might be better for analyzing data from epidemiologic studies and clinical trials.

The estimated regression coefficients obtained from the standard regression technique or from population-averaged models were the average value of individual regression lines. In contrast, the estimated regression coefficients obtained from multilevel modeling or subject-

specific models were the values of individuals [10, 22]. The question, then, is as follows: which method is preferable for analyzing longitudinal complex survey data? Although no concrete answer to this question is known, if the research question of interest is about a group of subjects, then the SR-RV estimation technique is appropriate, but if the research question of interest is about individual development, then MM-SW technique is appropriate [38].

In longitudinal data, there are two types of covariates: time-dependent (e.g., body mass index) and time-independent (e.g., sex). The standard errors of regression coefficients of time-dependent covariates can be underestimated and the standard errors of regression coefficients of time-independent covariates can be overestimated if the dependency among repeated measurements of each individual are ignored [11, 14]. Repeated measurements were treated as level 1 units that were nested within subjects, and subjects were treated as level 2 units that were nested within primary sampling units (PSU) in the NPHS. In the analyses of the NPHS data using the MM-SW technique, no correlation structure was used to take into account within-subject correlations. It is important to check whether the dependency of observations within a subject was considered precisely without considering the within-subject correlation structure. The estimated standard errors of the regression coefficients obtained from the analyses of the NPHS data using the MM-SW technique were larger than the SR-RV estimation technique. The reason for the larger standard errors as well as the wider 95% confidence intervals might be that the MM-SW technique did not accurately take into account the dependency among repeated measurements of each subject and the effects of sampling design. In contrast, SR-RV estimation technique might take into account within-subject correlation using an appropriate correlation structure and the effects of the sampling design, such as clustering and stratification, to estimate

the standard errors of the regression coefficients accurately using bootstrap variance estimation technique.

In order to analyze longitudinal complex survey data using the MM-SW technique, the weight variable for each level of complex survey data and the corresponding identification number are required. In real-life, publicly available complex survey data, it is difficult to obtain the corresponding identification number of sampling units because of privacy concern. So, it may not possible to determine the weight variable for each sampling unit for each level of real life complex survey data without knowing the identification number of each sampling unit. Many statistical software programs are available to analyze complex survey data, but not all statistical software can handle the weight variable for each level of complex survey data. Although both the MM-SW technique and the SR-RV estimation technique were used to analyze the longitudinal complex survey data, it is important to be cautious about using software to apply the MM-SW technique with dichotomous outcomes because the theory of multilevel modeling–scaled weights has not yet been developed as a universal feature in all statistical software programs.

The conclusion based on the analysis of longitudinal complex survey data is that the standard regression–robust variance (SR-RV) estimation technique might be the appropriate statistical technique compared with multilevel modeling–scaled weights (MM-SW) technique.

6.4 Objective 3: To investigate which statistical technique is optimal for analyzing cross-sectional complex survey data sets using a Monte Carlo simulation Technique

The third objective of this thesis was to assess and compare the performance of two statistical techniques: (i) the MM-SW technique and (ii) the SR-RV estimation technique by

analyzing simulated cross-sectional complex survey data via a Monte Carlo simulation technique.

The Monte Carlo simulation technique is an alternative to analytical methods. It is an empirical method based on random sampling from a known population to assess the behavior of a statistic. It is often impractical to collect data multiple times to assess the performance of the statistical technique. The main purpose of a Monte Carlo simulation technique is to produce artificial random samples multiple times (number of simulations or replications) from a known population and then to analyze these multiple random samples to investigate the behavior of a statistical procedure or methods of interest based on obtained results.

In Monte Carlo simulation technique, the RANTBL function in SAS® software program was used to generate 100 and 1000 cross-sectional complex survey data sets and, the sampling design of each simulated data sets were similar to the Saskatchewan data with sample size 1,731. The Saskatchewan data which is part of CHHS is a cross-sectional complex survey data and the sampling design of Saskatchewan data is similar to CHHS.

Both the MM-SW technique and the SR-RV estimation technique were applied to each of the 100 and 1000 simulated data sets, and the performance of both statistical techniques was assessed based on the assessment criteria: (i) bias of regression coefficients (ii) standardized bias of regression coefficients (iii) percentage bias of regression coefficients (iv) length of 95% confidence intervals of regression coefficients (v) coverage of true regression coefficients in the corresponding 95% C.I. obtained from simulated data and (vi) relative efficiency. These assessment criteria were estimated based on analysis of 100 and 1000 simulated data sets using both MM-SW technique and SR-RV estimation technique. The parameters estimates obtained from the analysis of 1000 simulated data sets were efficient and consistent based on estimated

values of assessment criteria compared to the analysis of 100 simulated data sets using MM-SW . The parameters estimates obtained from the analysis of 1000 simulated data sets were efficient and consistent based on estimated values of assessment criteria compared to the analysis of 100 simulated data sets using SR-RV estimation technique. Although the estimated bias for one covariates (education level) is lower for 100 simulated data sets compared to 1000 simulated data sets using SR-RV estimation technique.

Standardize bias and percentage biases are lower for this covariate due to the lower bias. Parameter estimates might be differed for multiple covariates in a model compared to single covariate in a model . The further study is required to figure out the reasons for this differences. Two covariates (BMI and education) were used in the model but no interaction was considered between these covariates using simulated data. It might be one of the reasons for the differences.

The overall simulated results indicated that the parameter estimates were efficient and consistent based on the analysis of data with the higher numbers of replications. The higher number of replications increased the accuracy and reliability of parameter estimates [100] which is the agreement of my obtained simulations results. The study in literature indicated that 1000 numbers of replications might be the reasonable sample size to obtain the reliable parameter estimates [100].

Based on the analysis of 1000 simulated data sets, the estimated bias, percentage bias, standardize bias for regression coefficients were higher in MM-SW technique for body mass index (< 25 kg/m^2, ≥ 25 kg/m^2) compared to SR-RV estimation technique but the estimated values of these criteria were almost similar between MM-SW technique and SR-RV estimation technique for education level (<secondary, $\geq$ secondary). The estimated biases of regression coefficient for body mass index (<25 kg/m^2, ≥ 25 kg/m^2) in MM-SW technique and

SR-RV estimation technique were 0.0082 and 0.0009 respectively (Table 5.13). These both estimated biases were small. These small values of biases indicated that both statistical techniques provided almost unbiased regression coefficients. The calculations of the percentage bias and the standardize bias were depended on bias. So, the values of percentage bias and the standardize bias were also differing between these two techniques but not much.

The estimated biases for regression coefficient of education level (<secondary, ≥ secondary) in MM-SW technique and SR-RV estimation technique were 0.0256 and 0.033 respectively (Table 5.13). These estimated biases were almost similar. So, the percentage bias and the standardize bias were also similar between these two techniques based on simulated data.

The results based on the calculation of bias, percentage bias and standardized bias for regression coefficients indicated that both MM-SW technique and SR-RV estimation technique provide unbiased regression coefficients in the simulation technique. The means square errors (MSE), length of 95% confidence intervals of regression coefficients and relative efficiency for both covariates were similar between MM-SW technique and SR-RV estimation technique (Table 5.13). The results from these assessment criteria indicated that the performance of both statistical techniques were comparable. The coverage rate of the true regression coefficients in corresponding simulated 95% confidence intervals were 91% for body mass index (< 25 kg/m^2, ≥25 kg/m^2) and 89% for education level (<secondary, ≥ secondary) in MM-SW technique. In contrast, the coverage of the true regression coefficients in corresponding simulated 95% confidence intervals were 83% for body mass index (<25 kg/m^2, ≥25 kg/m^2) and 81% for education level (<secondary, ≥ secondary) in SR-RV estimation technique. The results from

this assessment criterion also indicated that the performance was not differed between MM-SW technique and SR-RV estimation technique.

Several studies indicated that estimated regression coefficients and their standard errors were higher for the MM-SW technique than with the SR-RV estimation technique [6, 21, 24]. The results based on our simulation technique support the previously published results. The parameter estimates and their standard errors obtained from the simulation technique with 1000 simulated data sets provide the higher regression coefficients and their standard errors in MM-SW compared to SR-RV estimation technique except only the regression coefficients of education level (Table 5.13).

Based on the first objective, we found inconsistent results between the MM-SW technique and the SR-RV estimation technique. The estimated regression coefficients obtained from these two techniques were not consistent, but the standard errors and 95% C.I. were consistently smaller for the MM-SW technique compared with the SR-RV estimation technique. The difference in the results could be due to many things. First, cross-sectional and longitudinal complex surveys (longitudinal surveys were not the focus of my Monte Carlo simulation study) often have problems with missing values. Second, complex surveys based on stratification and clustering quite often have small sample sizes for some clusters. Simulated datasets based on the Monte Carlo or some other techniques do not have these problems. It is generally preferable to use a simulation technique with single or two covariates. Therefore, simulated data for a cross-sectional complex survey with only two covariates was generated. The results based on assessment criteria in Monte Carlo simulation suggested that there might not have huge difference of performance between the MM-SW technique and the SR-RV estimation technique to analyze the cross-sectional complex survey data.

The theoretical assumption of the SR-RV estimation technique was not similar to the theoretical assumption of the MM-SW technique. Because observations were assumed to be independent in the SR-RV estimation technique, the observations were assumed to be dependent within clusters in the MM-SW technique [6, 49, 50, and 98]. Standard regression with a binary outcome underestimates the standard errors of regression coefficients when it violates the assumption of independence [50].

The regression coefficients and standard errors were affected by the multilevel modeling technique if the sampling design is unbalanced [12]. In the standard regression–robust variance estimation technique, the bootstrap re-sampling variance estimation technique was used to estimate the standard errors, which takes into account the effect of design features of complex surveys and weight adjustments. The effects of the design features and the weight adjustments were taken into account using sandwich estimators in multilevel modeling – scaled weights technique. The design features such as stratifications, clustering and unequal probability of selection was taken into account in both techniques but the ways were different.

The conclusion based on the Monte Carlo simulation study is that the MM-SW technique and the SR-RV estimation technique are equally acceptable for analyzing cross-sectional complex survey data set. However, we observed low coverage which indicates there is a room for improvement in both methods for analyzing complex survey data.

6.5 Strengths

i) **Data sets:**

Both surveys (NPHS and CHHS) were conducted in all Canadian provinces and the sample sizes for both data sets were large. The power of statistical analysis was increased due to

large sample sizes. The longitudinal NPHS provides repeated measurements on each individual over time which further enhance the power of statistical analysis. Statistical analyses using NPHS and CHHS were conducted based on weighted data. Detailed information on risk factors is available in both data sets. The NPHS being longitudinal survey also provides information on time-dependent variables. Appropriates weights (overall weight to obtain the regression coefficient estimates and bootstrap weights to obtain the standard errors of regression estimates) were used to analyze CHHS and NPHS data sets. Hence, the results obtained from both surveys can be generalized to the entire Canadian population.

ii) Analytical Technique:

Both statistical methods used to analyze NPHS and CHHS data sets provided valid estimates of regression coefficients and their standard errors because non-response and design effects were taken into account by both methods at the analyses stages. Monte Carlo simulation technique is most commonly used either to compare more than two statistical methods or to identify the most appropriate or optimum statistical method to analyze a given data set obtained using a certain study design. Monte Carlo simulation technique was one of the main strengths of this thesis that was used to compare the performance of MM-SW technique and SR-RV estimation technique and to identify the preferable statistical technique to analyze complex survey data set.

6.6 Limitations

i) Data sets:

In both data sets (NPHS, CHHS), presence of type 2 diabetes was based on the positive response to the question "Do you have any of the following long-term conditions that have been

diagnosed by a health professional? – Diabetes". Hence the diagnostic criteria were self-reported physician diagnosed type 2 diabetes. In population health study, self-reported disease information is commonly used to measure the health status. The self-reported health information can be affected by gender, ethnicity and education. So, it is required to consider the validity of self-reported disease status. Validity of self-reported type 2 diabetes has been examined and reported by other researchers [105]. In CHHS data analyses, participants from Novas Scotia are not included in this analysis because the location of residence (rural, urban) was measured.

ii) Analytical Technique:

Theoretically, weight variables for each level are required to conduct the multilevel modeling analyses. However, overall weight generally is available for complex surveys conducted by Statistics Canada or statistical organizations of other countries. There is no literature available which recommends how to conduct multilevel modeling in the presence of missing weight information at each level. Therefore, as suggested by Rabe-Hesketh et al [6], weight of 1(one) was used at PSU and strata level. In the analyses of CHHS and NPHS based on multilevel modeling technique, might have led to unreliable results.

There are several methodological limitations especially in MM-SW such as: (i) multilevel pseudo maximum likelihood function is very complicated. The integral function is complex and it is not easily integrable. The numerical methods such as Newton-Raphson or Gauss-Hermite quadrature are commonly used to estimate parameters. (ii) Multilevel modeling assumes that covariates are measured at different levels but information on covariates at PSU and STRATA levels were not available in both data sets; (iii) Researchers proposed several scaling of weight methods but no one has recommended the best methods for scaling that can be used as a gold standard method; (iv) in practice, we often conduct the subgroup analysis (for

example: population >18 years and nine province out of 10 province for this thesis) but no literature is available that describes the adjustment of weights for subgroup analyses. Therefore, the weight variables which were available for both data sets for the entire Canadian population were used to conduct the analyses.

6.7 Future studies and recommendations

Further research is still needed to test the utility of MM-SW and SR-RV estimation techniques to analyze complex survey data, especially for longitudinal complex survey data. Some of the other areas related to complex survey data analyses which need attention are handling missing data and goodness of fit (especially for dichotomous outcomes). These areas are well developed for classical cross-sectional and longitudinal studies. But have received little attention for complex surveys.

Missing data is one of the important issues to be considered when analyzing complex survey data [1, 25]. The assumptions regarding missing data were different for the two techniques utilized in this thesis. The marginal model using GEE assumed the missing data were missing completely at random (MCAR), or in other words cases with complete data are indistinguishable from cases with incomplete data. In contrast, the multilevel model assumed the missing data was missing at random (MAR), or in other words the probability of the missing value depends on the observed variable [21, 62]. This might be the reason for the different results between the MM-SW technique and the SR-RV estimation technique. The effect of missing values on parameter estimation can be a new research area. As we found that the weight variable for each level of multistage complex survey is not commonly available in publicly used data sets or data sets available at Statistics Canada Research Data Centers. It would be an important area

to explore how missing weights information or missing data influence the results. Computation of weight variable for each level from overall weight variable and incorporating these weight variables for missing data will be an interesting area for future research.

Multilevel modeling does not have provision to specify various types of covariance structures in order to account for within-subject correlation to analyze the longitudinal complex survey data. It will be an interesting research area to explore how to incorporate the covariance structure in multilevel modeling procedure. Assessment of model fit and model diagnostics are not developed yet which can be a challenging work for future. Analysis of subsample data (obtained from complex survey) is commonly conducted to test special hypothesis and answer research questions but how to handle information related to design variables (strata, psu and weight) is not commonly known. Hence, the analysis of subsample data from complex survey will be an important future research area.

Statistical methodology depends on research question to analyze the complex survey data. If the research question or hypotheses is related to determine the impact on the population then SR-RV estimation might be the appropriate method. In contrast the MM-SW might be the appropriate method if the research question is related to determine the impact on an individual level. MM-SW might be the appropriate method if the weight variables are available in each level. SR-RV estimation might be the appropriate method if only overall weight or single level weight and bootstrap weights are available with the complex survey data. If the cluster size is small then SR-RV estimation technique might be the appropriate method compared to MM-SW.

BIBILIOGRAPHY

1. Fitzmore, G.M., Laird, N.M., Ware, J.H., Applied longitudinal analysis. A Jhon Wiley & Sons INC.Publication, 2004.

2. Korn, E.L., Graubard, B.I., Analysis of Health Surveys. A Wiley - Inter-Science Publication, John Wiley & Sons, 1999.

3. Molina, E.A., Smith, T.M.F. and Sugden, R.A., Modelling overdispersion for complex survey data. International Statistical Review, 2001. **69**(3): p. 373-384.

4. Demnati, A. and Rao., J.N.K., Linearization variance estimators for survey data. Survey Methodology, 2004. **30(1)**: p. 17-26.

5. Feder, M. Nathan, G. and Pfeffermann, D., Multilevel modeling of complex survey longitudinal data with time varying random effects. Survey Methodology, 2000. **26(1)**: p. . 53-65.

6. Rabe-Hesketh, S. and Skrondal, A., Multilevel modelling of complex survey data. Journal of the Royal Statistical Society Series a-Statistics in Society, 2006. **169**: p. 805-827.

7. Rao, J.N.K., Bootstrap methods for analyzing complex survey data. Proceedings of Statistics Canada Symposium, 2006.

8. Rao, J.N.K., Scott, A.J. and Skinner, C.J. , Quasi-score tests with survey data. Statistica Sinica, 1998. **8**(4): p. 1059-1070.

9. Steel , F., Multilevels models for longitudinal data. J.R. Statist.Soc. A., 2008. **171(1)**: p. 5-19.

10. Neuhaus, J.M., Kalbfleisch, J.D. and Hauck, W.W., A comparison of cluster-specific and population-average approches for analyzing correlated data. International Statistical Review, 1991. **59**(1): p. 25-35.

11. Twisk, J.W.R., Longitudinal data analysis. A comparison between generalized estimating equations and random coefficient analysis. European Journal of Epidemiology, 2004. **19**(8): p. 769-776.

12. Moerbeek, M., van Breukelen, G.J.P. and Berger, M.P., A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. Journal of Clinical Epidemiology, 2003. **56**(4): p. 341-350.

13. Mauny, F., Viel, J.F., Handschumacher, P. and Sellin, B., Multilevel modeling and malaria: a new method for an old disease. International Journal of Epidemiology 2004. **33**: p. 1337-1344.

14. Zeger, S.L. and Liang, K.Y. , Longitudinal data-analysis for discrete and continuous outcomes. Biometrics, 1986. **42**(1): p. 121-130.

15. Mohan V, Mathur, P., Deepa R., Deepa M., Shukla D.K., Menon G.R., Anand K., Desai N.G., Joshi P.P., Mahanta J., Thankappan K.R., Shah B, Urban rural differences in prevalence of self-reported diabetes in India. The WHO-ICMR Indian NCD risk factor surveillance, 2008. **80**: p. 159-168.

16. Pohar S.L., Majumdar, S.R., Johnson J.A., Health care costs and mortality for Canadian urban and rural patients with diabetes: population-based trends from 1993-2001. Clin. Ther. , 2007. **29**: p. 1316-24.

17. Kelly, C., Booth, G., Diabetes in Canadian Women. BMC Women's Health, 2004.

18. Nelder, J.A., and Wedderburn, R.W.M., Generalized linear models. J. R. Statist. Soc. A, 1972. **135(2)**: p. 370-384.

19. LaVange, L.M., Koch, G.G., and Schwartz, T.A., Applying sample survey methods to clinical trials data. Statistics in Medicine, 2001. **20**(17-18): p. 2609-2623.

20. Xue, X., Gange, S.J., Zhong, Y., Burk, R.D., Minkoff, H., Massad, L.S., Watts, D.H., Kuniholm, M.H., Anastos,K., Levine, A.M., Fazzari, M., D'Souza,G.Plankey, M., Plaefsky, J.M., Strickler, H.D., Marginal and Mixed-Effects Models in the Analysis of Human Papillomavirus Natural History Data. Cancer Epidemiology Biomarkers & Prevention 2010. **19**(1): p. 159-169.

21. Carrière, I., and Bouyer, J. , Choosing marginal or random-effects models for longitudinal binary responses: application to self-reported disability among older persons. Bmc Medical Research Methodology, 2002. **2**: p. 1-10.

22. Hu, F. B., Goldberg, J., Hedeker , D, Flay B. R., Pentz M. A., Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. American Journal of Epidemiology, 1998. **147(7)**: p. 694-703.

23. Rodriguez, G. and Goldman, N., An assessment of estimation procedures for multilevel models with binary responses. Journal of the Royal Statistical Society Series a-Statistics in Society, 1995. **158**: p. 73-89.

24. Kuchibhatla, M., Fillenbaum, G.G., Comparison of methods for analyzing longitudinal binary outcomes. Aging and Mental Health, 2003. **7(6)**: p. 462-468.

25. Rabe-Hesketh, S., and Skrondal, A., Multilevel and longitudinal modeling using stata. A Stata Press Publication, 2008.

26. Masaoud, E. and Stryhn, H., A simulation study to assess statistical methods for binary repeated measures data. Prev Vet Med. **93**(2-3): p. 81-97.

27. Hosmer, D.W., Lemeshow, S., Applied Logistic Regression. Wiley-interscience, John Wiley & Sons, Inc, 1989.

28. Rao, J.N.K., and Wu, C.F.J., Resampling inference with complex survey data. Journal of the American Statistical Association, 1988. **83**(401): p. 231-241.

29. Stata Survey Data Reference Manual 9. A Stata Press Publication, 2005.

30. Binder, D.A., On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 1983. **51**: p. 279-292.

31. Freedman, D.A., On the so-called "Huber Sandwich Estimator" and "Robust Standard Errors". American Statistician, 2006. **60**(4): p. 299-302.

32. Krewski, D. and Rao, J.N.K., Inference from stratified samples-properties of the linearization, jackknife and balanced repeated replication methods. Annals of Statistics, 1981. **9**(5): p. 1010-1019.

33. Efron, B., 1977 Rietzlecture - bootstrap methods - another look at the jackknife. Annals of Statistics, 1979. **7**(1): p. 1-26.

34. Yeo, D., Mantel, H. and Liu, T.P., Bootstrap variance estimation for the national population health survey. In Proceedings of the Survey Research Methods Section, American Statistical Association, 1999.

35. Mach, L., Saidi, A., and Pettapiece, R., Study of the properties of the Rao-Wu bootstrap variance estimator: what happens when assumptions do not hold? Proceedings of the Survey Methods Section, 2007.

36. Kovačević, M.S., Rong, H. and You, Y., Bootstrapping for variance estimation in multi-level models fitted to survey data. ASA section on Survey Research Methods.

37. Wedderburn, R.W.M., Quasi-likelihood functions, generalized linear-models, and gauss-newton method. Biometrika, 1974. **61**(3): p. 439-447.

38. Diggle, P., Heagetry P., Liang, K.Y., Zeger, S.L., Analysis of Longitudinal Data. Oxford Statistical Science Sereies 25, 2002. **Second Eidition**.

39. Zeger, S.L., Liang, K-Y., and Albert, P.S., Models for longitudinal data- a generaliazed estimating equation approach. Biometrics, 1988. **44**(4): p. 1049-1060.

40. Nelder, J.A., Quasi-likelihood and pseudo-likelihood are not the same thing. Journal of Applied Statistics, 2000. **27**: p. 1007-1011.

41. Hendricks, S.A., Wassell JT., Collins J.W., Sedlak S.L., Power determination for geographically clustered data using generalized estimating equations. Statistics in Medicine, 1996. **15**(17-18): p. 1951-1960.

42. Skinner, C. and Vieira, M.D., Variance estimation in the analysis of clustered longitudinal survey data. Survey Methodology, 2007. **33**(1): p. 3-12.

43. Rao, J.N.K., Marginal models for repeated observations: inference with survey data. Proceedings of the section on Survey Research Methods of the American Statistical Association, 1998: p. 76-82.

44. Crouchley, R. and Davies, R.B., A comparison of population average and random-effect models for the analysis of longitudinal count data with base-line information. Journal of the Royal Statistical Society Series a-Statistics in Society, 1999. **162**: p. 331-347.

45. Pahwa, P., McDuffie, H.H., and Dosman, J.A., Longitudinal changes in prevalence of respiratory symptoms among Canadian grain elevator workers. Chest, 2006. **129**(6): p. 1605-1613.

46. Leung, D.H.Y., Wang, Y.G. and Min, Z., Efficient parameter estimation in longitudinal data analysis using a hybrid GEE method. Biostatistics, 2009. **10**(3): p. 436-445.

47. Ren, Q. and Roberts, G., Marginal logistic regression models for longitudinal complex survey data. Proceedings of the Survey Methods Section, SSC Annual meeting, 2005.

48. Ghosh, S., Pahwa, P. Rennie, D.C., A comparison of design-based and model-based methods to estimate the variance using national population health survey data. Model assisted statistics and applications, 2008. **3**: p. 33-42.

49. Goldstein, H., Multilevel modeling of survey data. The Statistician 1991. **40**: p. 235-244.

50. Guo, G. and Zhao, H., Multilevel modeling for binary data. Annu.Rev. social, 2000. **26**: p. 441-462.

51. Goldstein, H., Multilevel statistical models. Edward Arnold, London, 1995.

52. Goldstein, H., Nonlinear multilevel models, with an application to discrete response data. Biometrika, 1991. **78**(1): p. 45-51.

53. Goldstein, H., Multilevel mixed linear-model analysis using iterative generalized least-squires. Biometrika, 1986. **73**(1): p. 43-56.

54. Goldstein, H. and Rasbash, J., Improved approximations for multilevel models with binary responses. Journal of the Royal Statistical Society Series a-Statistics in Society, 1996. **159**: p. 505-513.

55. Kovačević, M.S. and Rai, S.N., A pseudo maximum likelihood approach to multilevel modelling of survey data. Communications in Statistics-Theory and Methods, 2003. **32**(1): p. 103-121.

56. Asparouhov, T.M., Muthen, B., Muthen & Muthen, Multilevel modeling of complex survey data. Proceedings of the Joint Statistical Meetings(JSM) in Seattle. ASA Section on Survey Research Methods, 2006: p. 2718 - 2726.

57. Asparouhov, T., General multi-level modeling with sampling weights. Communications in Statistics-Theory and Methods, 2006. **35**(3): p. 439-460.

58. Grilli, L. and Pratesi., M., Weighted estimation in multilevel ordinal and binary models in the presence of informative sampling designs. Survey Methodology, 2004. **30(1)**: p. 93-103.

59. Pfefferman, D., Skinner, C.J., Holmes, D.J., Goldstein,H., Rasbash, J., Weighting for unequal selection probabilities in multilevel models. J.R.Statist.Soc. B, 1998. **60(1)**: p. 23-40.

60. Carle, A.C., Fitting multilevel models in complex survey data with design weights: Recommendations. Bmc Medical Research Methodology, 2009. **9**.

61. Graubard, B.I. and E.L. Korn, Modelling the sampling design in the analysis of health surveys. Stat Methods Med Res, 1996. **5**(3): p. 263-81.

62. Burton, A., Altman, D.G., Royston, P., P. Holder, R.L. , The design of simulation studies in medical statistics. Statist. Med, 2006. **25**: p. 4279-4292.

63. White, J.S., Greenland, S., Simulation study of hierarchical regression. Statistics in Medicine, 1996. **15**: p. 1161-1170.

64. Borroni, C.G., Luscia, F., A simulation study about robust methods for regression analysis. International journal of statistics and systems **1(2)**: p. 155-166.

65. VanNess P.H, Holford,T.R., and Dubin, J.A., Power simulation for categorical data using the RANTBL function. SUGI, Statistics and Data Analysis. **30**: p. 207-230.

66. Moineddin, R., Matheson, F.I., and Glazier, R.H., A simulation study of sample size for multilevel logistic regression models. Bmc Medical Research Methodology, 2007. **7**:34.

67. Fan, X., Felsővályi,Á. Sivo, S.A., Keenan, S.C., SAS® for Monte Carlo Studies: A Guide for Quantative Researchers,. SAS Institute Inc., , 2002.

68. Ng, E.S., Carpenter, J.R., Goldsten, H., and Rasbash, J., Estimation in generalized linear mixed models with binary outcomes by simulated maximum likelihood. Statistical Modelling, 2006. **6**: p. 23-42.

69. Gentle, J.E., Statistics and computing: Random number generation and Monte Carlo methods. Springer, 2003.

70. King, H., Aubert,R.E. and Herman,W.H., Global burden of diabetes, 1995-2025 - Prevalence, numerical estimates, and projections. Diabetes Care, 1998. **21**(9): p. 1414-1431.

71. Crispim, D., Canani, L.H., Gross J.L., Tshiedel B, Souto KEP, Roisenberg I. , Familial History of Type 2 Diabetes in Patientsw from southern Brazil and its Influence on the Clinical Characteristics of this Disease. Arq Bras Endocrinol Metab, 2006. **50**: p. 862-868.

72. Al-Moosa, S., Allin, S., Jemiai, N., Al-Lawati, J., Mossialos, E., Diabetes and urbanization in the Omani population: an analysis of national survey data. Population Health Metrics, 2006. **4**(5): p. (24 April 2006).

73. Bener, A., Zirie, M and Al-Rikabi, A., Genetics, obesity, and environmental risk factors associated with type 2 diabetes. Croatian Medical Journal, 2005. **46**(2): p. 302-307.

74. Azimi-Nezhad M, Ghayour-Moberhan M., Parizadeh MR, Safarian M, Esmaeili H, Parizadeh SMJ, Khodaee G, Hosseini J, Abasalti Z, Hassankhani B, Ferns G. , Prevalence of type 2 diabetes mellitus in Iran and Its relationship with gender, urbanization, education, marital status and occupation. . Singapore Med. J. , 2008. **49(7)**: p. 571-576.

75. Maty S.C., Everson-Rose.S.A., Haan M.N., Raghunathan T.E., Kaplan G.A. , Education, income, occupation, and the 34-year incidence (1965-99) of Type 2 diabetes in the Alameda County Study. International Journal of Epidemiology 2005. **34**: p. 1274-1281.

76. Sadeghi M, Roohafza.H., Shirani S., Poormoghadas M, Kelishadi R, Baghaii A, Sarraf-Zadegan N., Diabetes and Associated Cardiovascular Risk Factors in Iran: The Isfahan Healthy Heart Programme Annals Academy of Medicine 2007. **36(3)**: p. 175-180.

77. Mainous III, A.G., King, D.E., Garr, D.R., Pearson,W.S., Race rural residence, and control of diabetes and hypertension. Annals of Family Medicine, 2004. **2**(6): p. 563-568.

78. Tang, M., Chen ,Y., Prevalence of Diabetes in Canadian Adults Aged 40 Years or Older. Diabetes Care, 2000. **23**: p. 1704-1705.

79. Yun, S. and Lee, Y. , Comparison of hierarchical and marginal likelihood estimators for binary outcomes. Computational Statistics & Data Analysis, 2004. **45**(3): p. 639-650.

80. Archer, K.J., Lemeshow, S., Goodness-of-fit test for a logistic regression model fitted using survey sample data. The Stata Journal, 2006. **6 Number 1**: p. pp. 97-105.

81. Liang, K.-Y.and .Zeger, S.L , , Longitudinal data analysis using generalized linear models. Biometrika, 1986. **73(1)**: p. 13-22.

82. Pan, W., Akaike's Infirmation Criterion in Generalized Estimating Equations. Biometrics, 2001. **57**: p. 120-125.

83. Edwards, A.C., Canadian Heart Health Database 1986-1992. Memorial university of newfoundland, Canada, 1997.

84. Canada, H., Canadian guidelines for body weight classification in adults. Health Canada publications Centre, 2003.

85. National population health survey household component, c.-l.c.d., Statistics Canada, 2005.

86. Canada, H., Canadian Guidelines for Body Weight Classification in Adults. Health Canada publications Centre, 2003.

87. Rabe-Hesketh, S., Skrondal, A., and Pickles, A., Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. Journal of Econometrics, 2005. **128**(2): p. 301-323.

88. Rabe-hesketh, S., Skrondal, A, and Pickles,A., Reliable estimation of generalized linear mixed models using adaptive quadrature. The Stata Journal, 2002. **2(1)**: p. 1-21.

89. Rabi, D.M., Edwards,A. L., Southern, D.A., Svenson, L. W., Sargious, P.M., Norton,P., Larsen, E.T., and Ghali, W.A., Association of socio-economic status with diabetes prevalence and utilization of diabetes care services. Bmc Health Services Research, 2006. **6**:124.

90. Tang, M., Chen, Y., and Krewski, D., Gender-related differences in the association between socioeconomic status and self-reported diabetes. International Journal of Epidemiology, 2003. **32**(3): p. 381-385.

91. Booth, G.L. and Hux, J.E., Relationship between avoidable hospitalizations for diabetes mellitus and income level. Archives of Internal Medicine, 2003. **163**(1): p. 101-106.

92. Volkers, A.C., Westert,G.P. and Schellevis, F.G., Health disparities by occupation, modified by education: a cross-sectional population study. Bmc Public Health, 2007. **7**:196.

93. Bruce, S., Prevalence and determinants of diabetes mellitus among the Metis of western Canada. American Journal of Human Biology, 2000. **12**(4): p. 542-551.

94. Jiang, Y., Jiang, Y., Chen, Y., and Mao, Y., The contribution of excess weight to prevalent diabetes in Canadian adults. Public Health, 2008. **122**(3): p. 271-276.

95. Jiang, Y.,Chen, Y., Manue, D. Morrison, H., Mao,Y., and Obesity Working Group, Quantifying the impact of obesity category on major chronic diseases in Canada. ScientificWorldJournal, 2007. **7**: p. 1211-21.

96. Plotnikoff, R.C., Taylor,L.M. Wilson, P.M., Coumeya, K.S., Sigal, R.J., Birkett, N., Svenson, L.W., Factors associated with physical activity in Canadian adults with diabetes. Medicine and Science in Sports and Exercise, 2006. **38**(8): p. 1526-1534.

97. Kriska, A.M., Saremi,A., Hanson, R.L., Bennett, P.H., Kobes, S., Williams, D.E., and Knowler,W.C. Physical activity, obesity, and the incidence of type 2 diabetes in a high-risk population. American Journal of Epidemiology, 2003. **158**(7): p. 669-675.

98. Maas, C.J.M. and Hox, J.J., The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Computational Statistics & Data Analysis, 2004. **46**(3): p. 427-440.

99. Burton, A, Altman, D.G., Royston, P, Holder, R.L., The design of simulation studies in medical statistics. Statistics in Medicine 2006. 25: p. 4279-4292.

100. Díaz-Emparanza, I., Is a small Monte Carlo analysis a good analysis? Checking the size, power and consistency of a simulation-based test*. Statistical papers 2003. 43: p. 567-577

101. Heeringa, S.G., West, B.T. and Berglund, P.A., Applied survey data analysis. A Chapman & Hall book, CRC Press Taylor & Francis Group 2010.

102. Nemes, S, Jonasson,M.,J., Genell, Anna., Steineck, G., Bias in odds by logistic regression modelling and sample size. BMC Medical Research Methodology 2009. 9:56.

103. Pahwa, P and Karunanyake, C.P., Modeling of longitudinal polytomous outcome from complex survey data - application to investigate a association between mental distress and non-malignant respiratory diseases. BMC Medical Research Methodology 2009. 9:84.

104. Binder, D.A., Roberts, G.R., Statistical inference in survey data analysis: where does the sample design fit in? [http://socserv.socsci.mcmaster.ca/rdc2003/binderoberts.pdf]

105. Rhodes et al, Accuracy of Administrative coding for type 2 diabetes in children, adolescents, and young adults. Diabetes Care 2007, 30(9).

106. Cho, E, Rimm, E.B., Stampfer, M.J., Willett, W.C., Hu, F.B., The impact of diabetes mellitus and prior myocardial infarction on mortality from all causes and from coronary heart disease in men. Journal of the American College of Cardiology 2002. 40 : p.954-960.

107 Mason, W.M., Wong, G.M., Entwistle, B., Contextual analysis through the multilevel linear model. Sociol. Methodol. 1983. 13: p72-103.