PREVALENCE AND INCIDENCE OF COLORECTAL CANCER IN RURAL SASKATCHEWAN:

AN APPLICATION OF GENERALIZED ESTIMATING EQUATIONS (GEE) AND SURVIVAL ANALYSIS

A Thesis Submitted to the

College of Graduate and Postdoctoral Studies

in Partial Fulfillment of the Requirements

for the degree of

Master of Science (Biostatistics)

in the School of Public Health

University of Saskatchewan

Saskatoon

By

Abubakari Ibrahim Watara

PERMISSION TO USE

In presenting this thesis in partial fulfilment 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 the 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 allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

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

Director of School of Public Health
Health Sciences Building E-Wing,
University of Saskatchewan
104 Clinic Place
Saskatoon, Saskatchewan S7N 2Z4
Canada

OR
Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

ABSTRACT

Although agricultural activity is recognized to be associated with colorectal cancer (CRC) prevalence and incidence, little is known regarding the true prevalence and incidence or CRC risk factors related to farming. This study determines the prevalence of CRC both at baseline and follow-up among farm and non-farm rural residents, estimates the incidence of CRC, and further identifies risk factors for CRC incidence and prevalence. Data from the Saskatchewan Rural Health Study (SRHS) was collected in 2009 and a four-year follow-up in 2014 through completed questionnaires from 8,261 individuals (level 1) nested within 4,624 households (level 2) at baseline and 4,867individuals within 2,797 households at follow-up. A modified version of Dillman's methods for mail and telephone surveys was used to maximize response rates both at baseline (42%) and follow-up (63%). The study sample consist of 5,599 individuals at baseline and 3,933 at follow who were 50 years or older.

Hierarchy in data was accounted for using the generalized estimating equations (GEE) approach using the exchangeable covariance structure and the within-cluster correlations due to the longitudinal design were also accounted for using the PROC GENMOD robust variance estimation. Multilevel marginal logistic regression models based on GEE were fitted to determine risk factors for CRC. To determine risk factors for the incidence of CRC, the Cox proportional hazards (PH) regression model was used. The prevalence of CRC decreased over time among rural farm residents (baseline: 3.1%; follow-up: 1.3%, $p<0.05$), however increased among rural non-farm residents (baseline: 1.4%; follow-up: 2.0%, $p>0.05$). Individuals who spent their first year of life in a farm had higher risk of developing CRC that their counterparts who did not OR = 1.64, $p<0.05$). A similar relationship was observed for individuals who ever lived on a farm at one point in life times (OR = 1.29, $p<0.05$). Occupational exposure to grain dust and radiation were significant ($p<0.05$) determinants of longitudinal changes in CRC prevalence after controlling for important confounders.

The crude incidence rate of 1.98 per 1,000 person-years (i.e. 27/13,632 total time under observation and at risk), resulting in a cumulative incidence of 0.8% during the study period (2010 - 2014). The results show that quadrant, use of natural gas, living on a farm in one's first year of life, exposure to oil/gas well fumes and radiation, increasing age, increasing BMI, and female gender predict CRC incidence. We conclude that prevalence, longitudinal changes in the CRC prevalence among farming and non-farming rural residents as well as the incidence of CRC appear to depend on a complex combination of individual and contextual factors.

ACKNOWLEDGEMENTS

In the name of Allah, the Compassionate, the Merciful. Praise is only for Allah, the Lord of the Universe, the All-Compassionate, the All-Merciful, the Master of the Day of Judgment. Thee alone we worship and to Thee alone we pray for help. Show us the straight way, the way of those whom Thou hast blessed; who have not incurred Thy wrath, nor gone astray. I am very grateful to the Almighty God for His grace, mercies, kindness, protection and numerous blessings that He bestowed on me throughout this period of my life. It was by His mercy that helped me to complete this thesis.

Words cannot describe my sincerest gratitude to my supervisor Dr. Punam Pahwa for her compassion, moral support and inspiration, patience, advice and “only high-quality and no room for mediocrity” philosophy that have inspired me to complete this thesis. A special thank you to her for introducing me to the concept of GEE which forms an important part of this thesis. May God bless her abundantly.

To my thesis committee members Drs. Shahid Ahmed and Shahedul Khan, I sincerely thank them for monitoring my work, making efforts in reading my thesis and providing me with timely and valuable comments which have been extremely helpful in the successful completion of this thesis. I would also like to thank Dr. Cindy Feng for agreeing to chair my thesis committee at the shortest notice.

I would also like to acknowledge the Canadian Centre for Health and Safety in Agriculture (CCHSA), Dr. James Dosman, the principal investigator of the CIHR funded-Saskatchewan Rural Health Study for providing me with the data used in the thesis.

Finally, I would like to thank my parents and the entire Watara family for their innate love and support, encouragement and prayers. I would like to also thank my dearest friends Abdulai “Ababa” Abdul Karim, Gazali Issah and their respective families, whose friendship increased my faith and also diffused many challenges I faced while preparing this thesis. May Allah unite us again in Jannatul-Firdaus.

TABLE OF CONTENTS

PERMISSION TO USE i

ABSTRACT ii

ACKNOWLEDGEMENTS iv

TABLE OF CONTENTS v

LIST OF TABLES ix

LIST OF FIGURES xi

LIST OF ABBREVIATIONS xii

CHAPTER 1 – INTRODUCTION 1

1.1 Rational 1

1.2 Objectives and Research Questions 6

CHAPTER 2 – LITERATURE REVIEW 7

2.1 Statistical Methods for analyzing CRC outcomes 8

2.1.1 Modeling CRC Prevalence using Cross-sectional Data 8

2.1.2 Modeling Longitudinal Changes in CRC Prevalence using Clustered Binary Data 9

2.1.2.1 Generalized Estimating Equations (GEE) 11

2.1.3 Modeling CRC Incidence Using Survival Analysis Techniques 15

2.2 Epidemiology of Colorectal Cancer (CRC) 22

2.2.1 Colorectal Cancer Prevalence 22

2.2.2. International burden of CRC 23

2.2.3 CRC Incidence in Canada 25

2.2.4 CRC Incidence in Saskatchewan 25

2.2.5 Etiology and Risk Factors of CRC 26

2.2.5.1 Non-modifiable Risk Factors/Non-sporadic CRC 27

2.2.5.1.1 Age 27

2.2.5.1.2 Ethnic and Racial Background 27

2.2.5.1.3 Personal History of Colorectal Polyps 28

2.2.5.1.4 Family History of Adenomatous polyps 28

2.2.5.1.5 Personal History of Inflammatory Bowel Disease (IBD) 29

2.2.5.2 Modifiable Risk Factors 29

2.2.5.2.1 Dietary habits 29

2.2.5.2.2 Cigarette smoking 31

2.2.5.2.3 Physical inactivity and obesity 31

2.2.5.2.4 Alcohol consumption 32

2.2.5.2.5 Occupation and Occupational Exposures 33

CHAPTER 3 – DATASET DESCRIPTION 35
3.1 Study design 35
3.2 Study population, Sampling and Recruitment 36
3.3 Data Collection 37
3.4 Theoretical Framework 38
3.5 Study Variables 38
3.5.1 Operational definitions 39
3.5.1.1 Contextual factors 39
3.5.1.2 Covariates 43
3.5.1.3 Outcome 44
CHAPTER 4 – METHODS 45
PART I 46
OUTCOME VARIABLE: DISCRETE/DICHOTOMOUS 46
4.1 Models for Longitudinal Data Analysis 47
4.2 The Univariate Quasi-Likelihood Function 48
4.3 Generalized Linear Models (GLMs) for Longitudinal Data Analysis 50
4.3.1 The Generalized Estimating Equations (GEE) Methodology 50
4.3.2 The Multivariate Quasi-Likelihood Function 53
4.4 Marginal models for Longitudinal Data Analysis 54
4.4.1 Marginal Models for Discrete Outcomes 55
4.5 Research Question One (1): Marginal Modeling Approach 56
4.5.1 Statistical Application: Research Question One (1) 56
4.5.1.1 Estimating Crude Prevalence of CRC 56
4.5.1.2 Estimating adjusted Prevalence of CRC 57
PART II 59
OUTCOME VARIABLE: TIME-TO-EVENT/SURVIVAL TIME 59
4.6 Models for Survival Data Analysis 60
4.6.1 Introduction 60
4.6.2 Survival Analysis Techniques for Cross-sectional Independent Data 61
4.6.2.1 Survival time 61
4.6.2.2 Censoring 62
4.6.2.3 Functions of Survival time (adopted and modified from [2] and [14]) 63
4.6.2.4 Hazard Ratio (HR) 64
4.6.2.5 The Cox Proportional Hazards (PH) Regression Model 65

4.6.2.5.1 The Discrete Proportional Hazards (PH) Model ... 67
4.6.2.5.2 Discrete-Time Survival Analysis (DTSA) and the Logit Link ... 70
4.6.2.5.3 Discrete-Time Survival Analysis (DTSA) and the Complementary Log-Log (C-Log-Log) Link ... 71
4.6.3 Objective Two (2): Survival Analysis ... 72
4.6.3.1 Statistical Application: Objective Two (2) ... 72
4.6.3.1.1 Crude Incidence of CRC ... 72
4.6.3.1.2 Adjusted Incidence of CRC ... 73
CHAPTER 5 – RESULTS ... 75
5.1 Introduction ... 75
5.2 Sample population ... 75
5.3 Descriptive Characteristics ... 75
5.3.1 Population characteristics by quadrant ... 83
5.3.2 Distribution of colorectal cancer (CRC) by residence status ... 84
5.3.3 Baseline and follow-up characteristics of participants by CRC status ... 85
5.4 Research Question One (1): Prevalence Analysis ... 92
5.4.1 Crude Prevalence Rate Calculation ... 92
5.4.2 Adjusted Prevalence of CRC using Marginal Modeling Approach ... 94
5.4.2.1 Marginal Modeling Approach for Cross-sectional Data ... 94
5.4.2.2 Marginal Modeling Approach for Longitudinal Changes in CRC Prevalence ... 102
5.5 Research Question Two (2): Incidence Analysis ... 107
5.5.1 Crude Incidence Calculation ... 107
5.5.2 Adjusted Incidence Analysis using Cox's Proportional Hazards Regression Model ... 108
5.6 Conclusion: Summarizing methods used for data analysis ... 113
CHAPTER 6 – DISCUSSION ... 114
6.1 Introduction ... 114
6.2 Prevalence ... 114
6.2.1 Risk factors for CRC prevalence and longitudinal changes in prevalence ... 115
6.2.1.1 Environmental/Contextual factors ... 115
6.2.1.2 Lifestyle/individual factors ... 116
6.2.1.3 Early life exposures ... 117
6.2.1.4 Occupational Exposures ... 117
6.2.1.5 Family history of Cancer ... 117
6.2.1.6 Other risk factors (Increasing age and gender) ... 118
6.3 Incidence ... 118

6.3.1 Risk factors for CRC Incidence 119
6.3.1.1 Environmental/contextual factors........ 119
6.3.1.2 Lifestyle/individual factors........ 119
6.3.1.3 Early life exposures 119
6.3.1.4 Occupational exposure........ 120
6.3.1.5 Family history of Cancer........ 120
6.3.1.6 Other risk factors (Increasing age and BMI; gender)........ 121
6.4 Study Strengths and limitations 122
6.4.1 Study Strengths 122
6.4.2 Study Limitations........ 123
6.5 Conclusion........ 124
REFERENCES **125**
Appendix A: The SRHS Baseline Questionnaire **134**
Appendix B – GLMs with random and systematic components with link functions **150**
Appendix C – Notations of Vectors and Matrices........ **151**
Appendix D – Computational definitions when T is continuous or discrete........ **154**
Appendix E – Rural municipalities located in the four study quadrants of the SRHS Study. **155**

LIST OF TABLES

Table 2. 1 Major North American studies assessing CRC prevalence in a general Adult Population 22

Table 3. 1 Household income adequacy level by income and size of household 40

Table 5. 1 Population Characteristics stratified of rural residents (≥ 50 years) by farm and non-farm residence status, Saskatchewan 78

Table 5. 2 Population characteristics by quadrant at baseline 83

Table 5. 3 Population characteristics by quadrant at follow-up 84

Table 5. 4 Location of home by CRC cases and non-cases at baseline 85

Table 5. 5 Population characteristics stratified by CRC status and survey 87

Table 5. 6 Population Characteristics for CRC Cases and Non-cases at Baseline and Follow-up 90

Table 5. 7 Prevalence of Colorectal Cancer Stratified by Geographic Location and Farm/Non-Farm Residence at Time-point 1 (Baseline) 93

Table 5. 8 Prevalence of Colorectal Cancer Stratified by Geographic Location and Farm/Non-Farm Residence at Time-point 2 (Follow-up). 93

Table 5. 9 Univariable Analysis of the Dependency of Colorectal Cancer on Contextual Factors, Individual Factors, and Covariates by Odds Ratio (OR), 95% CI, and P-value by Time-point 96

Table 5. 10 Multivariable Analysis of the Dependency of Colorectal Cancer on Contextual Factors, Individual Factors, and Covariates by Odds Ratio (OR), 95% CI, and P-value by Time-point 99

Table 5. 11 Predicting Probability of CRC at baseline 101

Table 5. 12 Longitudinal Determinants of CRC Prevalence by Odds Ratio (OR), 95% CI, and P-value 104

Table 5. 13 Predicting Probability of CRC for Longitudinal Changes 106

Table 5. 14 New CRC cases stratified by time-point/waves.. 108

Table 5. 15 Cox PH regression analysis of CRC incidence during 4 years of follow-up according to baseline characteristics.. 110

Table 5. 16 Summarizing methods used for data analysis... 113

LIST OF FIGURES

Figure 2. 1 Common regression models for analyzing cross-sectional data for CRC prevalence 9

Figure 2. 2 Incidence rates for CRC by province and gender, Canada 2017, age-standardized to the 2011 Canadian population.. 26

Figure 3. 1 Flowchart showing selection of the study sample and sample size.. 37

Figure 3. 2 Conceptual framework 38

Figure 4. 1 Statistical methods available for independent and dependent outcomes in longitudinal data analysis 48

Figure 4. 2 Illustrative representation of selecting new CRC cases in the SRHS data set 73

LIST OF ABBREVIATIONS

CRC	-	Colorectal cancer
SRHS	-	Saskatchewan Rural Health Study
PHF	-	Population Health Framework
FDRs	-	First degree relatives
GEE	-	Generalized Estimating Equation
GLMs	-	Generalized Linear Models
OLR	-	Ordinary Linear Regression
TTE	-	Time-to-event
PH	-	Cox's Proportional Hazard Regression Model
DTSA	-	Discrete-Time Survival Analysis
C-Log-Log		Complementary Log-Log Link
HR	-	Hazard rate
CI	-	Confidence interval
ASR	-	Age-standardized rate
ASIR	-	Age-standardized incidence rate
CCR	-	Canadian Cancer Registry
PTCRs	-	Provincial and Territorial Cancer Registries
SCA	-	Saskatchewan Cancer Agency
SRHS	-	Saskatchewan Rural Health Study
LDA	-	Longitudinal Data Analysis
PCG	-	Primary caregiver
SSA	-	Sub-Saharan Africa
NCIRS	-	National Cancer Incidence Reporting System
GLOBOCAN-		Global cancer incidence, mortality and prevalence
HNCC	-	Hereditary non-polyposis CRC
IBD	-	Inflammatory bowel disease
SEER	-	Surveillance, Epidemiology and End Results
SES	-	Socioeconomic status
NHIS	-	National Health Interview Survey
CD	-	Crohn disease
AGA	-	American Gastroenterological Association
RR	-	Relative Risk
NIH	-	National Institutes of Health
WCRF	-	World Cancer Research Fund
AICR	-	American Institute for Cancer Research
HMPS	-	Hereditary mixed polyposis syndrome
PCHs	-	Polycyclic Aromatic Hydrocarbons
IARC	-	International Association of Research on Cancer
HPV	-	Human Papillomavirus
BMI	-	Body Mass Index
EPIC	-	European Prospective Investigation into Cancer and Nutrition

CHAPTER 1 – INTRODUCTION

1.1 Rational

Regression analysis is an important statistical analysis with a versatile application, which can be used to analyze cross-sectional, longitudinal and survival [time-to-event (TTE)] data [1]. In cross-sectional data, for each of n experimental units, there is only one response variable, y_i $(i = 1,2, \ldots, n)$ with a $kx1$ vector of covariates denoted by $\boldsymbol{x}$,$[\boldsymbol{x} = (x_1, x_2, \ldots, x_k)$, thus allowing for the measurement of the response and covariates at the same time [2]. However, in longitudinal data, for each experimental unit, there are repeated observations y_{it} $(i = 1,2, \ldots, n; t = 1,2, \ldots, k)$ with a $k\text{x}1$ vector of covariates over time t, denoted by $\boldsymbol{x_{it}}$, allowing for the measurement of changes in the same experimental units over a period of time, usually at baseline and follow-ups [2]. In survival data, the n experimental units are followed over a specified period of time (e.g. in days, weeks, bi-weekly, years, etc.), with a primary focus on the time at which an event of interest occurs [2].

Statistical methods for analyzing cross-sectional data have been well-developed and are easily implemented through available commercial software such as SPSS[1], SAS[2], STATA[3], R[4], WESVAR[5], and SUDAAN[6]. Depending on one's objective(s)/research question(s) and type of outcome, there are different regression techniques [1]. Estimates obtained from regression models for cross-sectional data are computed via various statistical algorithms. For example, the computation of estimates of (i) linear regression for Gaussian outcomes are based on least squares; (ii) logistic regression for dichotomous outcomes are based on maximum likelihood theory; (iii) semi-parametric survival analysis for follow-up data are based on partial-

[1] IBM Corp. Released 2017. Version 25.0 (http://www.ibm.com)
[2] SAS Institute, Inc. Cary, NC, version 9.1.3 (http://www.sas.com/)
[3] STATA, Stata Corp LP, 1996 (http://www.stata.com)
[4] R Development Core Team (2008) (http://www.r-project.com)
[5] WESVAR, Westat Inc., 2006 (http://www.westat.com/)
[6] SUDAAN, Research Triangle Institute, 2005 (http://www.rti.org)

maximum likelihood [3-4]. In medical research, alternative techniques for modeling cross-sectional data have also been developed. Some of these include; log-binomial regression, a generalized linear model (GLM) with logarithmic link function [5-6] and a complementary log-log model with a link function of the form $\log(-\log(1-\pi))$ [7].

Relative to the analysis of cross-sectional data, the analysis of longitudinal data is more complex. Repeated observations for each experimental unit over time tend to be correlated with each other. One therefore needs to account for the within-subject correlation (depending on its magnitude) when analyzing longitudinal data. For example, longitudinal studies with repeated measures and hierarchy in data require accounting for two layers of complexity: first, hierarchy in data [accounted for by generalized estimating equations (GEE)] and second, within-subject correlation due to repeated measures over time (accounted for by robust variance estimation techniques such as jackknife) [3-4, 8]. Moreover, missing observations or incomplete data (over time) for an experimental unit are not unusual in longitudinal studies, as not all experimental units are usually available to be measured, either at baseline or follow-up surveys, thus resulting in unbalanced data patterns. Such missing observations are said to be missing completely at random (MCAR) when the "missingness" is not related to the observed data, missing at random (MAR) when the "missingness" is dependent on the observed data [9]. However, when the "missingness" depends on both the observed and unobserved data, it is said to be missing not at random (MNAR) [9].

Statistical methods for analyzing longitudinal data using GEE have been developed since the mid-80s and became available in commercial software in the late 90's. For example, the procedures GENMOD, GLIMMIX, and NLMIXED were incorporated into SAS in the late 90's while MIXED procedure was available earlier. These SAS procedures including the "*nlme*" and "*lme*" functions in R and SPLUS all implement longitudinal data analysis. Estimates obtained from regression models for longitudinal data are based on the multivariate quasi-likelihood theory [3]. For Gaussian data, statistical methods such as the random effects model (REM), with estimated parameters based on the restricted maximum likelihood or iterative maximum likelihood methods, have been developed [10]. However, for non-Gaussian data, additional information (resulting in nuisance parameters), together with the first two moments (mean and variance) is needed to determine the likelihood function [2]. In the presence of these nuisance parameters, estimating other model parameters using maximum likelihood methods is difficult [3].

The GEE approach, re-invented by Liang and Zeger [3-4], based on multivariate quasi-likelihood theory, is capable of handling complexities in longitudinal/clustered data. In GEE approach, a multivariate generalization of the quasi-likelihood theory, coefficients are based on quasi-likelihood, by assuming a functional form of the marginal distribution at every time point (i.e. y_{ij}), making it suitable for non-Gaussian outcomes [3]. One advantage of GEE is that estimated coefficients are asymptotically Gaussian, and the model is still robust even when the correlation structure is incorrectly specified [4].

Traditional regression methods cannot be used to analyze survival data. In survival data analysis, the primary outcome of interest is not just whether or not an event occurred (as in traditional regression models), but also when the event occurred. The inability of traditional regression methods such as linear and logistic regression techniques to incorporate both the event occurrence and the time of occurrence make them unsuitable for survival data analysis. In addition, traditional regression methods do not have the capacity to handle censoring, a unique situation of missing observations in survival analysis, where, experimental units; (i) do not experience the event of interest before the end of the study or dropped out before the event occurs (right-censoring) or (ii) have already experienced the event of interest at the time they joined the study (left-censoring) or (iii) the specific time the event occurred is not known, but the first and last time the experimental unit was measured with and without the event respectively are available (interval-censoring).

Statistical methods have been developed to analyze survival data. An important property of statistical methods for survival data is their ability to handle censoring, which is overlooked in traditional methods including two-sample t-test and analysis of variance (ANOVA) for comparing the survival times of two or more groups respectively, and linear regression models. Moreover, statistical methods for survival data analysis have a higher statistical power to detect significant effects than methods for analyzing binary responses such as logistic regression [11]. Kaplan-Meier estimation [12] methods are used to estimate and graph survival times for different groups, by estimating survival probabilities using a product-limit formula. The computation of estimates from semi-parametric survival analysis for survival data is based on the partial-maximum likelihood theory. The log-rank test [13] is used to compare two survival curves by testing the null hypothesis of equivalent survival curves. Alternatives tests to the log-rank test include the Wilcoxon (also called Breslow test in SPSS), Flemington-Harrington, Peto, Tarone-Ware, and stratified log-rank tests, which are all variations of the standard log-rank test statistic by applying different weights at a specified failure time [14]. In STATA, the function "*sts test*" computes descriptive statistics for Kaplan-Meier curves, the log-rank statistics as well as other alternatives to the log-rank statistic. Similar to linear regression models, regression models in survival analysis allow for the determination of the effects of covariates on survival time. The Cox proportional hazards (PH) model [15], based on the proportional hazards assumption or the accelerated failure time model (AFT) [16], based on regressing logarithmic survival time on a covariate (and can be extended to a multivariable model [17-18]), are two common methods used to test covariates in survival regression model. The decision to choose between these two models depends on whether or not the PH assumption or the assumption that survival times follow a parametric distribution in the Cox PH and the AFT models respectively is met [11].

The aforementioned methods are parametric statistical methods widely used to analyze a dichotomous outcome. In particular, dichotomous logistic regression may be used to determine the prevalence of disease outcomes such as CRC [6]. Using the GEE approach, longitudinal changes in prevalence of a disease are determined by accounting for within-subject correlation due to repeated measurements in a longitudinal design [3-4, 19-20]. Survival analysis techniques have been used to determine risk factors for the incidence of CRC [29-30].

Colorectal cancer (CRC) is a malignant tumor which starts in the cells of the colon or rectum and can spread or metastasize to nearby organs and other parts of the body [22]. Although in recent years, several advances in CRC research have been made in Canada, the prevalence and incidence of the disease have not been studied extensively in Saskatchewan. In fact, to the best of my knowledge, to date, I am not aware of any study in rural Saskatchewan investigating the prevalence and incidence of CRC. In 2017, CRC was the second most commonly diagnosed cancer in Canada (excluding non-melanoma skin cancers), accounting for about 26,800 incident cases, which represents 13% of all new cancer cases [23]. CRC incidence is higher amongst men than women in Canada [23]. In 2017, the age-standardized rate (ASR) of CRC was 79.6 and 54.9 per 100,000, with a 5-year net survival estimate of 63% and 65% in men and women respectively [23]. The reason for this unequal CRC statistics among men and women remains ill-defined in the literature. Current data show that the percentage of rural population in Saskatchewan is nearly two times the national proportion (Saskatchewan, 33% [24]; Canada, 18% [23]). The major occupation of these rural dwellers is farming. Further research is needed to examine the prevalence and incidence of CRC in rural Saskatchewan, by occupational exposure (farm and non-farm rural residents) using longitudinal data.

In this thesis, the Saskatchewan Rural Health Study (SRHS), a longitudinal survey dataset was used analyzed. The primary aim of this thesis is to determine the prevalence and incidence of CRC in rural Saskatchewan using the GEE and survival analysis techniques respectively.

The baseline survey of SRHS was conducted in 2010 with a follow-up in 2014.[7] The primary aim of the SRHS was to test the hypotheses that, individual level (smoking, alcohol consumption, obesity) and contextual (access to health services and socio-economic) factors are associated with adverse health outcomes (CRC). To date, the SRHS dataset has not been analyzed to determine the prevalence, longitudinal

[7] Refer to section Chapter 3 for a detailed description of the SRHS dataset

changes in prevalence as well as significant risk factors for the prevalence and incidence of CRC using appropriate statistical methods.

1.2 Objectives and Research Questions

Overall objective: To determine the prevalence and incidence of CRC among rural dwellers in Saskatchewan.

To address this objective, we will undertake the following specific questions:

Research questions

1. What is the prevalence of self-reported doctor-diagnosed CRC and associated risk factors in rural Saskatchewan using GEE and robust variance estimation techniques?
2. What is the incidence of self-reported doctor-diagnosed CRC in rural Saskatchewan using survival analysis techniques?

The remainder of this thesis is organized as follows: A review of relevant literature is presented in Chapter 2. In subsections 2.1, I present a literature review of existing methodology for analyzing CRC outcomes for prevalence, longitudinal changes in prevalence and incidence. In subsection 2.2, I present the epidemiology of CRC. I describe the SRHS dataset in chapter 3 while the statistical methods for data analysis are presented in Chapter 4. The results of this thesis are presented in Chapter 5. I provide a discussion and conclusion in Chapter 6.

CHAPTER 2 – LITERATURE REVIEW

In this chapter, I present a review of existing literature on the statistical methods for analyzing CRC outcomes for prevalence, changes in prevalence over time, and incidence. CRC data are often collected from cross-sectional, longitudinal and survival studies. The major difference among these is that cross-sectional data only allows for the determination of the covariate-outcome relationship at a single point in time while longitudinal data contain information of the same subjects, followed over a period of time extending beyond a single point in time [2]. As a result, observations from the different time points in the latter tend to be correlated with each other [2]. In survival analysis, subjects are followed over a period of time to determine if an event of interest occurs or not and also when it occurred. The outcome of interest in survival analysis is usually survival time [14]. Censoring usually occurs in survival data [14].

Depending on the data and the objective of one's research, appropriate models are applied. Statistical methods for longitudinal data must account for the within-subject correlation and/or hierarchy in data while methods for survival data analysis must take into consideration follow-up time and censored observations [14]. Ignoring these needs may result in severely biased estimates of model parameters, further resulting in deceptive inferences. Statistical models presented in this chapter are grouped for cross-sectional, longitudinal and survival data.

The remainder of this chapter is organized as follows: In section 2.1, I present statistical models for analyzing CRC outcomes. In section 2.1.1, I present a review of statistical methods for modeling CRC prevalence. In sections 2.1.2, a review of statistical methods for modeling longitudinal changes in prevalence and methods for modeling CRC incidence are presented in section 2.1.3. I present the epidemiology of CRC in section 2.2.

2.1 Statistical Methods for analyzing CRC outcomes

2.1.1 Modeling CRC Prevalence using Cross-sectional Data

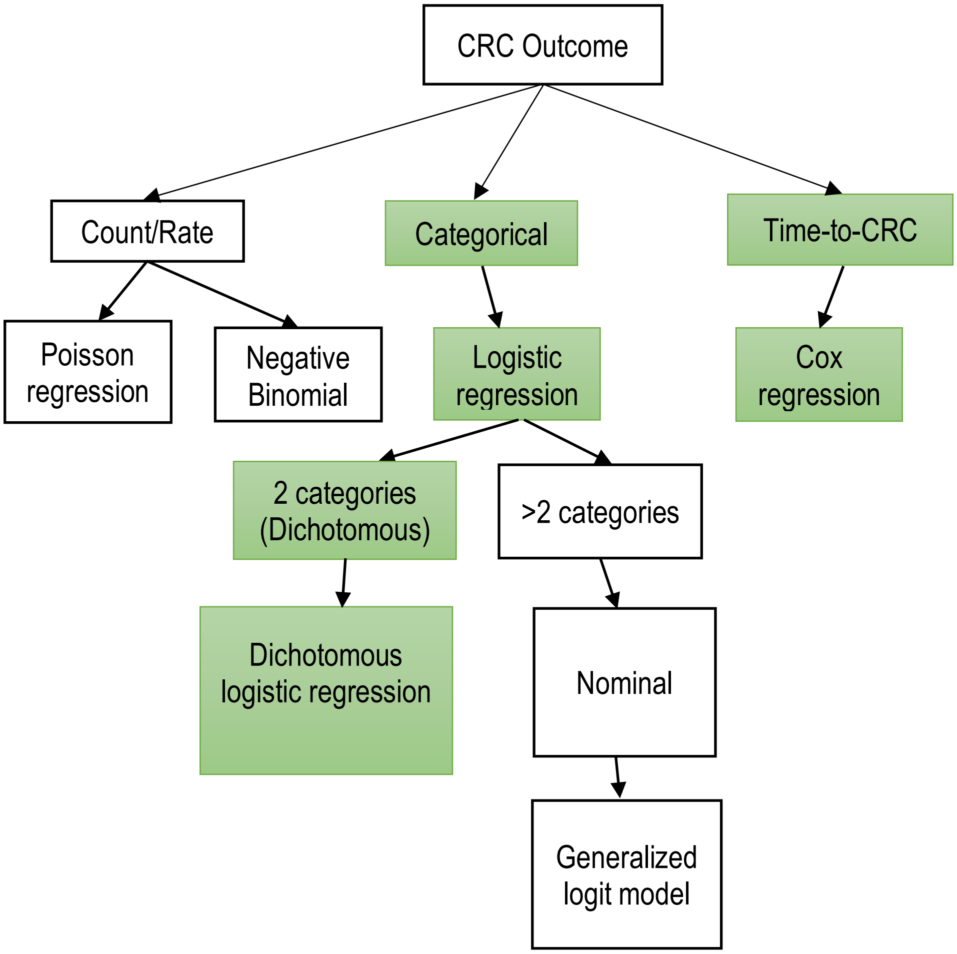
Regression models are commonly used to analyze cross-sectional data to determine significant risk factors for CRC prevalence. Depending on the nature of the CRC outcome of interest (i.e. categorical, count/rate, or time-to-CRC), different regression models are available. In the early 1980s, stratification and standardization were also used to estimate prevalence and prevalence rate ratios (PRR) of disease with categorical outcomes [25]. In cross-sectional data analysis, logistic regression is used to model the relationship between a single binary outcome and a set of covariates. In particular, dichotomous logistic regression is used to model binary outcomes with only two levels/categories (e.g. presence or absence of CRC) [26]. The odds ratio (OR) which represents the odds that an event will occur given a particular covariate, compared to the odds of the event occurring in the absence of that covariate, is used to measure the association between the outcome and a covariate [26].

The Cox's proportional hazards (PH) regression model [15], initially developed to estimate instantaneous conditional hazards ratio using censored could also be modified risk factors for the prevalence of disease outcomes in cross-sectional studies. In fact, Breslow [27] first suggested the use of Cox regression in analyzing cross-sectional data by imposing equal follow-up times for all experimental units. Subsequently, Lee and Chia [28] applied the Breslow's modification of the Cox model, called the Breslow-Cox model, in a cross-sectional study of 236 employees occupationally exposed to cadmium, a CRC-causing chemical, to estimate the prevalence of cadmium while adjusting for confounding and effect modifiers. The Breslow-Cox model can easily be implemented using various software.[2,8] The procedure PHREG can fit the Breslow-Cox model in SAS. In medical research, several alternative techniques for modeling cross-sectional studies are

[8] EPICURE, Hirosoft International Corp., (http://www.risksciences.com/project/epicure/)

discussed. Some of these include; log-binomial regression, a generalized linear model (GLM) with logarithmic link function and a complementary log-log model with link functions of the form $\log(-\log(1-\pi))$ [5-7]. Figure 2.1 summarizes some widely used regression models for analyzing cross-sectional data for CRC prevalence.

Figure 2. 1 Common regression models for analyzing cross-sectional data for CRC prevalence



2.1.2 Modeling Longitudinal Changes in CRC Prevalence using Clustered Binary Data

Research in longitudinal data analysis became popularized in the late 50s, with the advent of computational tools for statistical purposes [29]. Contributions from several authors including Box [30], Geisser et al [31], Rao [32-33], and Potthoff et al [34], and Grizzle and Allen [35] led to the development of a rich class of regression models for only Gaussian outcomes [29]. The models developed during this period were largely based on the traditional maximum likelihood theory [29]. However, for non-Gaussian outcomes, there was little work done before 1986 particularly for binary outcomes [2]. This was due, in part, to the lack of the multivariate Gaussian distribution for the joint probability distribution of y_{it}, the response for the i^{th} experimental unit at the t^{th} $(t = 1,2, \dots n_i)$ time [2]. Unlike in Gaussian longitudinal data analysis, in non-Gaussian or discrete longitudinal data analysis, the first two moments (i.e. mean and variance) are not sufficient to fully specify the likelihood function [2]. Hence additional information (resulting in many nuisance parameters) is required [36-38]. For example, recall that if random variables Y and Z have binomial distribution with parameters n and p and Poison with parameter λ respectively, then$E(Y) = np$, $var(Y) = np(1 - p)$ and $E(Z) = var(Z) = \lambda$. This functional interdependence of the mean and variance on each other makes it impossible to model them separately leading to difficulties in both computation and interpretation in non-Gaussian data analysis [29].

For the regression models developed prior to 1986, the traditional methods of maximizing the likelihood function could not be applied in this case because of the presence of too many nuisance parameters together with the regression coefficients [2]. Moreover, for non-Gaussian longitudinal data, the integral of the likelihood function does not have a closed-form, further making it impossible for the use of methods associated with its maximization [39]. Alternative likelihood methods were developed, but for only a few specific instances of non-Gaussian data. The marginal models developed by Bahadur [40], Bishop et al [41], and Fitzmaurice et al [38] are log-linear marginal models commonly used to formulate a probability model for multivariate binary data as a function of canonical parameters [41]. In particular, interpreting the canonical parameters in the model by specified by Bishop [41] was dependent on the number of responses. A fundamental challenge in using such log-linear models is that, since the number of responses often differ across subjects in a typical longitudinal study and may also increase rapidly with time, evaluating the likelihood function becomes problematic. To unify the regression techniques discussed above, thus overcoming the complexities associated with longitudinal/clustered data analysis, Liang and Zeger [3] proposed the generalized estimating equations (GEE) for Gaussian and non-Gaussian data alike, based on the multivariate quasi-likelihood theory. The GEE method is reviewed next.

2.1.2.1 Generalized Estimating Equations (GEE)

Liang and Zeger re-invented the GEE approach as a class of estimating equations for both Gaussian and non-Gaussian longitudinal data, based on the multivariate quasi-likelihood theory [3]. The quasi-likelihood theory was proposed by Wedderburn [42]. GEE as proposed by Liang and Zeger, is a multivariate generalization of the work of Wedderburn [3]. While the GEE approach by Liang and Zeger unified the regression models for various Gaussian and non-Gaussian responses [3], the generalized linear models (GLMs) proposed by McCullagh and Nelder unified the regression models for various Gaussian and non-Gaussian covariates regardless of the type of response variable [43]. In GLMs, the response variable y_i is assumed to follow an exponential family distribution such as linear, logistic, Poisson, Negative Binomial as well as some parametric survival regression models [43]. There are three components to GLMs; (1) a random component for identifying the response variable and its sampling distribution; (2) a systematic component for specifying the covariates X_1, X_2, X_3,…,X_k used in the linear predictor function $\boldsymbol{x}_i{}^T\boldsymbol{\beta}$ (defined as $\beta_0 + \beta_1\, x_1 + \beta_2\, x_2 + \beta_3\, x_3 + \cdots + \beta_k\, x_k$); and (3) a link function [η or *g*(μ)] for 'linking' the random and systematic components of the GLM [44]. The SAS procedure GENMOD fits GLMs and data analysts need to specify the appropriate link function in the *LINK* option and the distribution of the response in the *DIST* option (*bin* for binomial, *poi* for Poisson, *negbin* for negative binomial, and *mult* for multinomial distributions). Appendix B summarizes some common GLMs with all three components.

Liang and Zeger extended the univariate quasi-likelihood proposed by McCullagh and Nelder [45] to account for within-subject correlation due to the longitudinal design, using multivariate quasi-likelihood approaches [3]. A detailed discussion of quasi-likelihood methods is provided by elsewhere [42-45].

The GEE approach proposed by Liang and Zeger is a multivariate analog of the quasi-likelihood method [3]. GEE is proposed mainly for marginal modeling with GLM [3]. It 'avoids' the use of the multivariate Gaussian distribution but assumes a functional form for the marginal distribution of y_{it} at every time point, making it also suitable for non-Gaussian outcomes [3]. In the GEE approach, the correct specification of the marginal means, variance, and the link function connecting covariates and marginal means are required [3-4, 45]. One advantage of GEE is that solutions are consistent, and the model is still robust even when the correlation structure is incorrectly specified [4]. By "*robust",* Zeger and Liang showed that sample variance-covariance matrix produces a consistent estimate of the population variance-covariance matrix even when the "*working*" correlation matrix is incorrectly specified [4].

In the past few decades, GEE has been applied to determine changes in different CRC outcomes over a period of time, while taking into account within-subject correlation due to the longitudinal design. Stevens et al. [46] investigated the effect of fecal occult blood testing (FOBT) for CRC on alcohol consumption, physical activity, smoking and fruit and vegetable consumption from the year 2008 to 2015 using GEE approach. The effects of FOBT involvement, time and group-time interactions on these risk factors were determined while accounting for the within-group (FOBT participants vs. non-FOBT participants) correlation. Authors fitted five GEE models, each included two main effects for group and time, and a group-time interaction effect, and controlled for education, occupation, ethnicity, retirement status, baseline wave, and chronic illness. Group-time interaction was significant for smoking behaviors over time (OR = 1.15; 95% CI: 0.90 – 1.47) [46]. The proportion of FOBT participants consuming fruits and/or vegetables did not significantly change (OR = 1.32; 95% CI: 0.91 – 1.90) [46]. Multivariable logistic regression with baseline and follow-up variables showed that retirement status at follow-up was positively associated with FOBT participation for CRC (Odds ratio - OR = 1.99; 95% CI: 1.25 – 3.15) [46]. The proportion of current smokers and drinkers reduced over time from 19.9% to 12.7% (OR = 0.74; 95% CI: 0.62 – 0.89) and from 65.9% to 61.6% (OR: 0.69; 95% CI: 0.53 – 0.91) respectively [46]. Over time, participants meeting physical activity guidelines did not change significantly at baseline (88.6%) and follow-up (85.5%) (n = 736; OR = 0.75; 95% CI: 0.49 – 1.15) [46].

Rabeneck et al. [47] applied GEE methodology to determine factors significantly associated with bleeding and perforation among individuals aged 50 to 75 years who underwent an outpatient colonoscopy in Ontario, Nova Scotia, British Columbia, and Alberta, Canada from April 1, 2002, to March 31, 2003. Authors used GEE models, with individual patients as the unit of analysis and two binary outcome variables (presence or absence of bleeding and perforation), while controlling for patient's (age, gender, comorbidity based on hospital admission, having polypectomy) as well as endoscopic factors (endoscopic specialty and experience). For models using data from the four provinces, authors treated "province" as a cluster while for models which used data from only Ontario, "physician" was the cluster. Separate GEE models were fitted to determine risk factors for colonoscopy-related bleeding and perforation. Multivariable GEE models of patient factors showed that age (OR = 1.61; 95% CI: 1.20 – 2.16, $p = 0.001$), gender (male is reference category) (OR = 0.52; 95% CI: 0.36-0.74, $p = 0.0003$), polypectomy (OR = 10.32; 95% CI: 6.52 – 16.34, $p < 0.0001$) were significantly associated with colonoscopy-bleeding only except comorbidity score (OR = 1.69; 95% CI: 0.33 – 8.65, $p = 0.53$) [47] . All but gender (OR = 1.21; 95% CI: 0.97 – 1.50, $p = 0.09$) were significantly related to colonoscopy-related perforation ($p < 0.05$) [47]. GEE models predicting bleeding or perforation showed that all patient factors were significant ($p < 0.05$) [47]. In Ontario, all endoscopic factors were significantly associated with bleeding and perforation ($p < 0.05$) [47]. Based on the individual GEE models, male gender, older age, having a polypectomy, and undertaking colonoscopy done by a low-volume endoscopist were significantly associated with higher odds of colonoscopy-related bleeding or perforation [47].

Glanz et al. [48] used the GEE approach to identify significant determinants associated with the intention to undertake genetic counseling and CRC susceptibility testing among Caucasian, Japanese and Hawaiian ethnicities. Data collected on respondents included background information, familial history of cancer, interest in seeking counseling on genetic colon cancer and intention to undertake genetic testing for colon cancer. The outcome variables are dichotomous (interest in seeking genetic counseling and intention to undertake genetic testing for colon cancer). GEE approach was used to identify significant covariates by accounting for within-cluster correlations. The SAS procedure GENMOD to was used to calculate generalized models using GEE. After adjusting for covariates, family intraclass correlations of 0.06 and 0.09 were reported for individuals who had interest in going for counseling on genetic colon cancer as well as those who had the intention of undertaking colon cancer susceptibility testing respectively. About 45% and 26% of respondents declared an interest in seeking genetic counseling and intended testing for colon cancer respectively [48]. Binary logistic GEE regression models showed that, education (some college: OR = 1.76; 95% CI: 0.95 – 2.27, *p* = 0.07; ≥college graduate: OR = 1.92; 95% CI: 1.02 – 3.60, *p* = 0.04), Hawaiian ethnicity (OR = 2.68; 95% CI: 0.91 – 7.89, *p* = 0.07), cancer worry (OR = 2.94; 95% CI: 1.61 – 5.37; *p* = 0.001), and family support (colon cancer in first degree relative – FDRs, OR = 1.12; 95% CI: 0.63 – 2.02; total cancers in FDRs, OR = 1.05; 95% CI: 0.78 – 1.41) were significantly associated with individuals who showed interest in seeking counselling for on genetic colon cancer [48]. Cancer worry (OR = 1.99; 95% CI: 1.16 – 3.42, *p* = 0.01), risk perception (OR = 1.84; 95% CI: 1.29 – 2.62, *p* = 0.001), and older age (OR = 1.28; 95% CI: 1.01 – 1.60, *p* = 0.04) were positively, and Japanese ethnicity (OR = 0.36; 95% CI: 0.17 – 0.79, *p* = 0.01) was inversely related to the intention to undertake colon cancer susceptibility testing [48].

In a prospective cohort study, Stürmer et al. [49] determined the risk of CRC after starting the use of nonsteroidal anti-inflammatory drugs (NSAIDs) in over 20,000 male physicians age 40 to 84 years from 1982 to 1988 using the GEE method. GEE approach was used to estimate the propensity for regular (>60 days/year) NSAID use [49]. To determine covariates associated with regular NSAID use, logistic regression using GEE was used to account for an individual's correlated use of NSAID at different time points [49]. Regular NSAID use increase from 2 to 56% over the period [49]. Logistic GEE regression models revealed that significant predictors of NSAID use included body mass index (BMI), age, medication use, alcohol consumption, gastrointestinal diseases, coronary artery disease (CAD), headache, arthritis, and hypertension ($p < 0.05$) [49]. In addition, at least five years of regular use of NSAID was associated with a relative risk (RR) for CRC of 1.0 (95% CI: 0.7 – 1.5) [49].

2.1.3 Modeling CRC Incidence Using Survival Analysis Techniques

The standard regression techniques described in sections 2.1.1 and 2.1.2 cannot be used to analyze CRC survival data. This is because these methods do not take into consideration follow-up time and censored observations [14]. As a result, survival analysis techniques, which primarily studies the distribution of survival times (i.e. time to an event) as well as whether the event occurs or not, have been developed to overcome the limitations of traditional regression methods, which only studies the latter scenario [14].

No one knows when the concept of survival analysis was birthed, but it is believed to be one of the oldest statistical methods [50]. In the original stages of the transformation of survival analysis, the life-table methodology was used to estimate survival functions from subjects' lifetimes with delayed entry or left truncation (i.e. when a subject is not observed at baseline but only from a subsequent follow-up) and right censoring (i.e. when the subject leaves the study before the occurrence of the event) [50]. Survival analysis did not appear to be integrated with statistics at the time [50]. As a result, several authors including Westergaard [11] highlighted major drawbacks in the life-table methodology including the fact that it did not account for sampling variability. In the 1950's, the life-table methodology was presented to the statistical and medical community in very seminal surveys including that of Ederer and Cutler [52]. In these surveys, time was measured in discrete units [52]. As a result, survival frequencies from time t to $(t + 1)$ are multiplied together to form the survival probability across these time periods [52]. However, there is a limitation to this approach in that, developing approximations based on the discrete grouping of time (which is actually continuous) is very difficult [50]. To eliminate the need for such approximations, the Kaplan-Meier estimation (K-M) technique [12], a non-parametric method based on the work of Bohmer [53] was developed in 1958. In K-M estimation, survival, as well as censored times, are known [12]. In a way, Kaplan and Meier [12] formalized the concept of delayed entry by adjusting for the risk set, the set of experimental units alive and are being measured at a specific point of the time variable in question.

In the notation of Anderson and Keiding [50], let $(\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_n)$ be right-censored sample of observation times for $(X_1, X_2, \dots, X_n)$ respectively - the true survival times. Assume that $X_i, i = 1,2 \dots, n$ are independent and identically distributed (iid) nonnegative random variables with a common survival function $S(t)$. Right-censoring is often common in survival data [50]. So, the only available information for a specific subject i is a censoring time U_i (i.e. time lapsed by surviving till U_i without the event of interest occurring). More generally, any sample survival data includes $[(\tilde{X}_i, D_i), i = 1,2, \dots, n]$, where $\tilde{X}_i = \min(X_i, U_i)$ and D_i is the indicator, $I(X_i \leq U_i)$ of being uncensored [50]. Under the assumption of independent or non-informative censoring [i.e. for any subject in the risk set (with exceptionally high or low risk of failing) the probability of being censored at time t should not depend on the subject's prognosis for failure at time $X_i > t$], the survival distribution function $S(t)$ is estimated by the Kaplan-Meier estimator (K-M), denoted $\widehat{S(t)}$. Mathematically,

$$\widehat{S(t)} = \prod_{\tilde{X}_i \leq t} \left[1 - \frac{D_i}{Y(\tilde{X}_i)}\right] \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (2.1)$$

where $Y(t) = \sum I(\tilde{X}_i \geq t)$ is the number of subjects at risk just before time t. The K-M estimator in equation (2.1) is a non-parametric maximum likelihood estimator (NPML) and is approximately normally distributed with mean $S(t)$ and variance $\widehat{\sigma^2(t)}$ [50]. Peterson [54] showed that the K-M estimator, $\widehat{S(t)}$ is a consistent estimator of $S(t)$. Greenwood [55] proposed a formula for estimating the variance of the K-M estimator $\sigma^2(t)$, called the Greenwood's formula using the delta-method as:

$$\widehat{\sigma^2(t)} = \left[\widehat{S(t)}\right]^2 \left[\sum\nolimits_{\tilde{X}_i \le t} \frac{D_i}{Y(\tilde{X}_i)[Y(\tilde{X}_i)-1]}\right] \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots . (2.2)$$

Nelson [56-57] proposed an alternative non-parametric method, the cumulative survival function, denoted for estimating the survival distribution function $S(t)$. To estimate $S(t)$ with the new method, one may use the Nelson-Aalen estimator, which is defined as follow:

$$\widehat{A(t)} = \sum_{\tilde{X}_i \le t} \frac{D_i}{Y(\tilde{X}_i)} \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots .. (2.3)$$

Peterson [54] published the relationship between $\widehat{S(t)}$ and $\widehat{A(t)}$ as $\widehat{A(t)} = -\log\,(\widehat{S(t)}$. Although survival probabilities computed with K-M estimator are more interpretable, those obtained with the Nelson-Aalen estimator are more generalizable to multistate cases (i.e. modeling two or more life events simultaneously) beyond the traditional context of survival data [50].

To compare the survival curves or distribution of two or more independent groups, Mantel [58] proposed the log-rank test. To date, it is the most commonly used non-parametric test used in this regard and it tests the null hypothesis that two more survival curves are the same [14, 50]. One advantage of the log-rank test is that we do not need to know the probability distribution or shape of the survival times [50]. So, in a way, the log-rank test is the survival analog of the Mann-Whitney U test of means. The above techniques discussed are univariate because they describe survival based on one covariate but ignore the impact of others [50]. To account for the impact of other covariates that may affect survival time but were left out in the univariate survival analysis, regression models for survival data analysis become useful [15]. In 1972, Cox [15] extended the K-M life table by developing a multivariable survival regression model, called the Cox's Proportional Hazards (PH) regression model to account for the impact of several covariates on survival, based on the assumption that, the effects of covariates on survival are constant over time.

Cox's PH model relates a hazard function, an instantaneous rate of failure of an individual at time t, with a given specification of a vector of p predictor variables, $\boldsymbol{X_i} = (X_1, X_2, \dots, X_p)'$ to a baseline hazard $h_0(t)$ through;

$$\lambda_i(t, X_i) = \lambda_0(t)\exp(\boldsymbol{\beta' X_i}), \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.4)$$

where $\lambda_0(t)$ is the baseline hazard and $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is a vector of regression coefficients. Notice that when $X_i = 0\ (i = 1,2,\dots,p)$, $\lambda_i(t, X_i) = \lambda_0(t)$, This shows that, the baseline hazard describes a common distribution of survival time for all subjects while $\exp(\boldsymbol{\beta' X_i})$, the relative risk (RR) function describes an individual subject's hazard. The RR component of the PH model in equation (2.4) above is specified completely and so is modeled using parametric methods while the baseline hazard is unspecified, hence modeled using non-parametric methods [14]. This, therefore, makes the Cox's PH model semiparametric (i.e. partly parametric). Given this semi-parametricity, inferences on the likelihood for β are drawn using partial likelihood theory [50]. In fact, Cox [15] provided a partial likelihood function for β, called the Cox's partial likelihood as;

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n}\left[\frac{\exp(\boldsymbol{\beta' X_i})}{\sum_{j\in R_i}\exp(\boldsymbol{\beta' X_j})}\right]^{D_i} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.5)$$

where $R_i = \{j : \tilde{X}_j \geq \tilde{X}_i\}$ is the risk set of all subjects still alive and uncensored at time $\tilde{X}_i$ and D_i is an indicator of not being censored. When β is equal to its maximum partial likelihood estimator (MPL), $\boldsymbol{\hat{\beta}}$, the cumulative baseline hazard denoted $\widehat{A_0(t)}$ using the Breslow estimator defined as;

$$\widehat{A_0(t)} = \sum_{\tilde{X}_i \leq t}\frac{D_i}{\sum_{j\in R_i}\exp(\boldsymbol{\widehat{\beta'} X_j})}. \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (2.6)$$

Since Cox [15] published the PH model, several extensions have been made to it. In instances where covariates are time-dependent, a modified Cox's partial likelihood is obtained by replacing $\exp(\boldsymbol{\beta' X_j})$ in equation (2.5) with $\exp[\boldsymbol{\beta' X_j}(\boldsymbol{\tilde{X}_i})]$ [50]. However, modifying the Breslow estimator to account for the changing covariates is more complicated than simple replacement because, Cox's equation does not model the extra randomness arising from the time-dependent covariates, as a result, model estimates are not directly interpretable [50]. In such situations, the use of a joint model for the time-dependent covariate and the hazard using joint modeling data techniques [59].

Several studies have applied survival analysis techniques to identify risk or prognostic factors for CRC outcomes. Rasouli et al. [60] to determine significant prognostic factors and their impact on CRC survival in Iran. In the univariate analysis, K-M estimation method and log-rank tests were used to determine survival among various subgroups for 335 CRC cases. Cox's Proportional hazards model was used to determine the effect of CRC risk factors on survival. Authors reported a one- and five-year survival rate of 87% and 33% respectively [60]. In the Cox's multivariable analysis, age at diagnosis (65 years or older) (HR = 2.08; 95% CI: 1.17 – 3.71), being a worker (HR = 2.09; 95% CI: 1.22-3.58), wealthy economic status (HR = 0.51; 95% CI; 0.31 – 0.82), educational level (≤ Diploma) (HR = 0.61; 95% CI: 0.39 – 0.92), single patients (HR = 1.62; 95% CI: 1.10 – 2.40), and poor differentiation of tumor (HR = 2.25; 95% CI: 1.37 – 3.69) were significantly associated with CRC survival [60]. The log-rank test revealed no significant association between CRC survival and family history, sex, histology type, and tumor site ($p > 0.05$) [60]. However, age at diagnosis ($p < 0.001$), marital status ($p < 0.001$), place of residence (city vs. village) ($p = 0.0023$), level of education ($p < 0.001$), occupation ($p = 0.003$), comorbidity ($p = 0.026$), economic status ($p = 0.002$), smoking ($p = 0.017$), and tumor grade ($p < 0.001$) were significantly associated with CRC survival [60]. Wang et al. [21] reported higher CRC survival rate (HR = 0.94; 95% CI: 0.89 – 0.99) among persons 40- to 50-years in the univariate Cox's Proportional Hazards regression model, while higher rates are reported elsewhere among patients younger than 40 years [61].

In Malaysia, Hassan et al used survival analysis techniques to identify prognostic factors for CRC among 1,214 patients using data from the Malaysian Cancer Patient Registry – Colorectal Cancer [62]. Diagnosed CRC cases confirmed by histology were included in the study. Demographic information (i.e. age ethnicity, gender, family history of CRC, diabetes status, and current status of patient – alive or dead), pathological features (e.g. tumor size, lymph nodes, and stage at diagnosis) and treatment modalities were analyzed. Three- and five-year survival rates were computed using the Kaplan-Meier method. The log-rank test was used in the univariate analysis of all variables as well as in comparing the survival rates of each variable. Significant variables were further analyzed using multiple Cox proportional hazards regression to determine the hazard ratio of associated factors on CRC-survival and the final model was adjusted for gender, age, and ethnicity. Variables with p-values less than 5% were considered statistically significant. The overall three- and five-year survival rates were 59.1% (95% CI: 56.4% - 61.9%) and 48.7% (95% CI: 45.8% - 51.7%) respectively [62]. In the univariate analysis, primary tumor size ($p < 0.001$), stage at diagnosis ($p < 0.001$), treatment modalities ($p = 0.001$), and involvement of lymph nodes ($p < 0.001$) were significant predictors of CRC-survival [62]. No socio-demographic variable was found significant [62]. In the multivariable Cox regression analysis, primary tumor size ($p < 0.001$), stage at diagnosis ($p < 0.001$), treatment modalities ($p < 0.001$), and involvement of lymph nodes ($p < 0.001$) were significant prognostic factors for CRC-survival after adjusting for gender age and ethnicity [62]. Patients who only had a surgical resection as cancer treatment had the highest risk of dying from CRC as compared to those who underwent other methods of treatments (hazard ratio – HR = 1.71; 95% CI: 1.23 – 2.37) [62]. Overall, Malaysian patients with localized tumors had better CRC prognosis as compared to those with the disease at advanced stages [62].

In a similar study, Yuan et al. [63] determined prognostic factors among 837 Chinese CRC patients between 1996 and 2006 using survival analysis techniques. Univariate analyses and Cox proportional hazard regression models were used to identify significant prognostic factors for CRC. Patients were followed at three-month intervals for first 2 years, and six-month for the 3rd-5th year and the median follow-up was 45 months [63]. K-M method was used to plot survival curves while the log-rank test was used to compare differences in survival curves. Cox proportional hazard analysis was used to compute univariate hazard ratios for the identification of significant and independent prognostic factors for CRC-survival. The stepwise procedure was set to a 5% threshold while a p-value of > 0.05 defined statistical significance [63]. Three-year and five-year survival rates were reported at 74% and 68% respectively [63]. In the univariate analysis, age, tumor node metastasis, invasion of adjacent organs, lymphovascular invasion, tumor size, pathological type, the status of resection, histological grade, serum carcinogenic antigen level of diagnosis, preoperative obstruction were significant prognostic factors for CRC (p-value < 0.05) [63]. In the multivariable analysis, depth of bowel invasion, histological grade, and the number of metastatic lymph nodes significantly influenced CRC survival (p-value < 0.05) [63]. In particular, lymph node ratio (total positive lymph nodes divided by the total number of lymph nodes) was a strong predictor of stage III CRC ($p < 0.0001$) [63].

Elsewhere, Sharkas et al [64] identified determinants of CRC in a Jordanian population from 2005 to 2010 and computed survival rates for the disease using survival analysis techniques. Similar to Hassan et al. [62] and Yuan et al. [63], Sharkas et al [64] estimated overall survival using K-M estimation method and also compared survival rates between groups using standard log-rank test. However, the authors reported 5-year and 10-year survival rates of 58.2% and 51.8% respectively [64]. Over the six-year follow-up period, the 5-year survival rate for the under-50 age cohort decreased from 60.4% to 49.3% ($p < 0.005$) [64]. In Cox's proportional hazards regression model, age, stage, grade and location of the tumor were significantly associated with CRC-survival ($p < 0.005$) [64]. In particular, increased age, advanced stage CRC, poor differentiation of tumor, and right-sided CRCs were significantly associated with lower CRC survival rates [64].

2.2 Epidemiology of Colorectal Cancer (CRC)

2.2.1 Colorectal Cancer Prevalence

Like in many parts of the world, the availability of studies in the literature regarding CRC prevalence in Canada is limited. A meta-analysis of selected North American studies by Heitman et al reveal that, CRC prevalence ranged from 0 to 1.68% [142]. Reports that assessed CRC prevalence were located (See Table 2.1).

Table 2. 1 Major North American studies assessing CRC prevalence in a general Adult Population

Study	Mean Age (Years)	Male (%)	CRC Prevalence	Year
Spellman et al [143][a]	NR	46	**0.48**	2007
Kim et al [144]	58.1	44.4	**0.13**	2007
Rex et al [145]	NR	62.5	**0.20**	1993
Stevens et al [146]	NR	49.8	**0.37**	2003
Prajapati et al [147]	62.0	40.5	**0.39**	2003
Johnson et al [148]	65.0	67.8	**1.10**	1990
Imperiale et al [149]	68.6	44.6	**0.70**	2004
DiSario et al [150]	NR	41.2	**1.68**	1991
Mehran et al [151]	NR	49.5	**1.10**	2003
Pickhardt et al [152]	57.8	59.0	**0.16**	2004

NR, not reported
[a] Abstract only

It appears there is a variation of CRC prevalence within and between countries and across studies due to: (1) geographic differences in the distribution of CRC, (2) differences in methods used to identify CRC, (3) differences in the definitions of CRC across studies, and (4) the absence of standardized instruments in diagnosing CRC, and (5) differences in target populations. Within most regions and countries, CRC prevalence increases with age and is higher among men than among women [152]. Low Age-specific prevalence rates of CRC are reported for persons aged 50-54 years old (Male: 0.08 – 0.36%; Female: 0.05

– 0.23%) while higher rates are reported for older age groups across the world [153]. For instance, a prevalence rate about 0.13% among men aged 70-74 in Africa and South-central Asia while rates ranging between 1.4 – 1.8% are reported in Australia and Europe [153]. CRC is more prevalent in males than among females with an overall male-to-female prevalence ratio of 1.3 in South America (Male: 0.31%; Female: 0.23%) as compared to a prevalence ratio of 1.9 in Southern Europe (Male: 0.84; Female: 0.44%) [153].

In 2017, greater variations were reported even within same geographical regions [153]. For instance, in Europe, an estimated overall CRC prevalence among men aged 50-74 years varied from 0.19% in Albania to about 1.01% in Slovenia [153]. Higher prevalence rates were reported among men in Czech Republic (1.13%), Slovakia (1.19%), and Hungary (1.27%), and among women, for Denmark (0.66%), Norway, and The Netherlands (0.64%, both) [153]. On the other hand, lower prevalence rates for CRC were recorded in Albania (Male: 0.19%; Female: 0.14%) [153].

2.2.2. International burden of CRC

CRC is the third most frequently diagnosed cancer (after prostate and lung cancers) in the world among males (663,000 cases, 10%) and second (after breast cancer) among females (570,000 cases, 9.4%) [65]. It has an average of more than 1 million cases occurring annually [65]. Globally, CRC accounts for more than 9% of total cancer incidence [65-66], with more than 1.4 million cases in 2012 [66-67]. The incidence of CRC is expected to grow by 80% in 2035, with cases increasing (from the 2012 figure) to about 2.4 million worldwide, resulting in about 1.3 million people losing their lives [68]. In 2017, nearly half (44.6%) of the world incidence of CRC occurred in Asia, with China singularly contributing 18.6% of the global CRC incidence [68]. The Republic of Korea contributed the highest ASRs of CRC worldwide for both sexes combined, further highlighting the impact of Asia on the global CRC burden [68].

Australia, Europe, New Zealand, and North America have the highest rates of CRC incidence in the world [69] while the lowest rates are recorded in Central America, Africa, and South-to-Central Asia [66].

There is evidence that some countries with historically low CRC incidence are currently experiencing increasing risks of the disease. For example, since the mid-90s, Thailand and Japan have been facing increasing CRC incidence [70], about the same time rates in the Philippines also started to increase [71]. Rates in Saudi Arabia have more than doubled since 1994 [72], and the incidence of CRC has been progressively increasing in Iran since 1980 [73].

Elsewhere to the East, CRC incidence is also reported to be gradually increasing in Jordan [74], Singapore, China, and South Korea [75]. The "Westernization" of many countries in the West Pacific Region (WPR) including the Philippines, China, Korea and Japan, shown by high prevalence of smoking, sedentary lifestyles, obesity, excessive meat consumption, and high calories intake are putative risk factors for the increasing risks of the disease in the region [76].

In Sub-Saharan Africa (SSA), CRC is relatively rare, accounting for only about 2-6% of overall cancers [77]. When combined, the crude incidence of CRC in SSA is reported to be about 4.04 per 100,000 persons (Males: 4.38; Females: 4.38) [77]. However, some countries within the region have rates lower/higher than these. For instance, in Zimbabwe and Gambia, the age-adjusted rates (per 100,000) varies from 1.5 and 2.5 to about 8.5 to 7.1 among men and women respectively [78]. For countries with historically low incidence within SSA, such as Nigeria and Ghana, CRC incidence has started increasing contributing about 10-50% of overall cancers [79-80].

CRC incidence rates are higher in men than in women [65-66] with an overall male-to-female ratio associated with the ASR of CRC being 1.4:1 for many countries [65]. The reason(s) for this variation is ill-defined, but probably show complicated interactions between gender-specific exposure to risk factors and protective effects of endogenous and exogenous hormones as well as gender-specific variation in attitude towards screening for the disease [81].

2.2.3 CRC Incidence in Canada

Currently, CRC is the second most common cancer diagnosed in Canada, accounting for about 26,800 (Male: 14,900; Female: 11,900) incident cases, representing 13% of all new cancer cases in 2017 [82]. Of this total, 9,400 Canadians, representing 12% of all cancer deaths, died from the disease. On average, about 73 Canadians were diagnosed with CRC every day in 2017 and of this number, 26 people died every day of the disease in the same year [82]. Nearly 1 in 13 Canadian men will develop CRC at a point in his lifetime and 1 in 29 will die of this disease while 1 in 16 Canadian women will develop the disease in her lifetime and 1 in 34 will die from same [82].

In the same year, the incidence rate for this disease was 79.6 and 54.9 per 100,000 men and women respectively [82]. For the period 1988 to 2010, the ASR of CRC decreased by 11.3% and 14.5% among men and women respectively. It is projected that the incidence of CRC in Canada will increase by 79% in 2030 [82].

2.2.4 CRC Incidence in Saskatchewan

In 2017, CRC was the second (i.e. after prostate cancer) most frequently diagnosed cancer among males, with 500 new cases, representing 17.85% of all cancers diagnosed in Saskatchewan; third (i.e. after breast and lung cancers) among females, also with 380 new cases representing 13.57% [83]. Figure 2.3 shows the ASIR (per 100,000) for CRC were 89.9 and 60.6 among men and women in Saskatchewan, respectively [82]. These were the second and third highest ASIR of the disease in Canada respectively, after that of Newfoundland and Labrador. Recent Saskatchewan data reveal that, the number of incident cases of CRC will increase from 880 in 2017 to 1120 in 2032, representing a 27.3% jump [82].

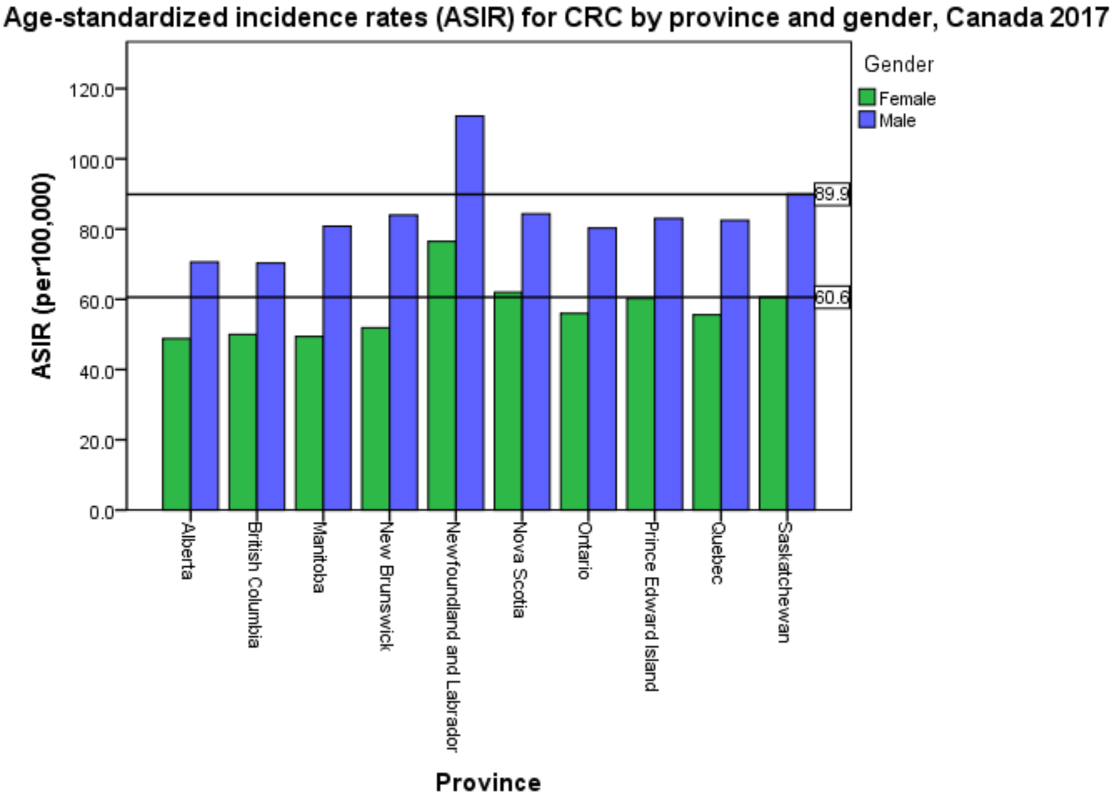


Figure 2. 2 Incidence rates for CRC by province and gender, Canada 2017, age-standardized to the 2011 Canadian population. Created by Watara using data from Canadian Cancer Registry (CCR) and National Cancer Incidence Reporting System (NCIRS) databases at Statistics Canada [123].

2.2.5 Etiology and Risk Factors of CRC

Colorectal cancer appears to be a multifactorial disease whose etiology remains ill-defined in the literature. The incidence and prevalence of the disease also appear to depend on a complex interplay between ecological factors and genetic vulnerability. So far, aside from age, gender, and race, which have been generally established as risk factors of the disease, other factors are inconsistently reported.

A large proportion (33%) of Saskatchewan residents are rural dwellers [24], which is higher than the national percentage (18%) [23]. A significant proportion of these rural dwellers in Saskatchewan are farmers. To the best of our knowledge, there is not enough information about the risk factors for CRC among farming and non-farming residents in rural Saskatchewan. Risk factors for CRC may be broadly categorized into two groups, namely non-modifiable and modifiable risk factors. The former majorly comprise of genetic factors while the former is made up of environmental factors. The following sections present current knowledge regarding these factors.

2.2.5.1 Non-modifiable Risk Factors/Non-sporadic CRC

2.2.5.1.1 Age

Age is a major risk factor for CRC. The risk associated with the disease increases with increasing age albeit it affects all ages [84]. However, CRC is rare among persons younger than 40 years of age [85] but increases significantly from 40 to 50 years, with age-specific incidence rates further increasing in each decade thereafter [85]. Over 90% of CRC cases are diagnosed persons aged 50 years or older [86]. CRC incidence is more than 50% higher among persons age 60-69 and 70-79 years than among persons younger than 40 years [86], even though some studies suggest an increasing incidence within the latter age cohort [87]. In fact, CRC is now none of the top 10 most frequently diagnosed cancers among individuals aged 20 to 49 years in the United States [88].

Gender differences vary by age. Recent Canadian data show that, over this period 2009-2013, CRC incidence for age groups 15-29, 30-49, 50-69, 70-84, and 85+, new cases of CRC increased by 4%, 8%, 11%, 15%, and 17% respectively, of all overall cancers in Canada [82]. For both genders combined, CRC was the most commonly diagnosed cancer (17%) in 2017, among persons aged 85 years and older in Canada [82]. More than half (54%) of all newly diagnosed CRC cases occurred individuals age 70 years or older in 2017 in Canada [82].

2.2.5.1.2 Ethnic and Racial Background

Variations in CRC incidence among different races/ethnicities have been reported in the literature. For example, using CRC data from the Surveillance, Epidemiology and End Results (SEER) program in the United States for the period 1975-2002, Irby et al. [89] a reported higher incidence of CRC among Blacks than Whites. It appears these differences depend on intricate interactions between screening and the predisposition to other etiologic factors, which, in turn, may depend on socioeconomic status (SES). In fact, the National Health Interview Survey (NHIS) in the United States reported that Whites had regular CRC screening than Blacks [90]. Moreover, Blacks had higher odds of developing CRC at a younger age of diagnosis and of the proximal type than Whites [89].

2.2.5.1.3 Personal History of Colorectal Polyps

Polyps are benign growths on the inner walls of either the colon or rectum. Colorectal polyps also called adenomatous polyps, are precursor lesions in the development of CRC [91]. Almost all (95%) of CRC cases result from these adenomas [92]. Individuals with a personal history of adenomatous polyps have a higher risk of developing CRC, especially with multiple polyps when compared to individuals with no such history [93-94]. Adenomas take about 5-10 years to become malignant [94]. Detecting and removing an adenoma before it becomes malignant is reported to lower CRC risk [95] even though complete removal of an adenomatous polyps increases the risk of developing subsequent (metachronous) colorectal cancer [95]. In particular, a personal history of large (greater than 1cm) polyps, for example, tubulovillous adenoma (TVA), elevates the risk of developing CRC, with a relative risk (RR) estimated between 3.5 to 6.5 [95-96]. On the other hand, individuals with a personal history of at most 2 small polyps (smaller than 1cm), do not have a significant increase in the risk of developing metachronous CRC [96].

2.2.5.1.4 Family History of Adenomatous polyps

More than 30% of individuals diagnosed with CRC have a past history of the disease in their family [97]. Individuals with first-degree relatives (such as a parent, child or sibling) who have had CRC in the past, have 2 to 4 times the risk of getting the disease when compared to individuals with no such family history, subject to the number of relatives affected and/or age at diagnosis [98]. In fact, individuals with first-degree relatives younger than 60 years or individuals with at least two first-degree relatives at any age who has been diagnosed of the disease, have elevated risks of developing CRC themselves [99]. There is emerging evidence that familial risks associated with CRC go beyond first and/or second relatives [100]. The reasons for this variation in risk are incompletely known but may be influenced by hereditary and environmental factors or a complex interaction of them.

2.2.5.1.5 Personal History of Inflammatory Bowel Disease (IBD)

Inflammatory Bowel Disease (IBD) is a long-term condition is which involves the inflammation of the colon and/or rectum. The risk of developing CRC among individuals with IBD is twice the risk among IBD-free individuals [101]. The two most common forms of IBD are ulcerative colitis (UC) and Crohn's disease (CD), even though, UC is better understood than CD [102]. UC and CD elevate an individual's risk of developing CRC [102]. In a meta-analysis of 116 studies, the total prevalence of CRC among UC patients is reported to be 3.7% (95% CI: 3.2 – 4.2) [103]. CRC risk increases with the degree, duration, and severity of the IBD [82-83]. The relative risk of developing CRC among individuals with IBD has been reported to be 4- to 20-fold [91] and further increase by 0.5 to 1.0% annually for 8 to 10 years after diagnosis [103].

2.2.5.2 Modifiable Risk Factors

2.2.5.2.1 Dietary habits

Diet as a risk factor for CRC has been extensively studied in the literature. Improvements in dietary habits can reduce about 70% of CRC incidence [104]. Dietary habits may directly influence CRC risk through various dietary elements such as calcium, vitamin D, etc. and indirectly through obesity and over-nutrition [91]. Variations in dietary fiber is believed to influence geographic differences in CRC incidence rates [91]. The WCRF and AICR, both have proposed that dietary variations are consequent for the differences in CRC rates between "Westernized" and African countries [105]. The intake of dietary fiber may increase stool volume and transit time which reduces CRC risk due to less exposure to carcinogens [105]. In an observational study, Negri et al [106] reported a negative relationship between the intake of dietary fiber and the risk of CRC [107]. However, recent randomized-control studies are inconclusive about whether or not dietary fiber is a risk factor for CRC [107]. Inadequate levels of vitamin D in the blood is also associated with an increased risk of developing CRC, albeit some studies are still inconclusive [107].

Studies assessing the potential association between the intake of fruits and vegetables and the risk of CRC have reported inconsistent results. In a meta-analysis evaluating the impact of several factors on CRC risk, a significant 15% reduction in CRC risk for 3 or more fruit servings per day was reported ($p = 0.02$) [108]. A similar conclusion was reached for vegetable consumption (RR = 0.86 per 5 servings/day) [108]. However, no association was observed between fruit/vegetable intake and CRC risk in a similar meta-analysis [109]. Currently, such inconsistent results are not completely understood and remain ill-defined in the literature.

Data implicating the consumption of red (e.g. beef, mutton, pork, goat, lamb/veal, horse) and/or processed meat (e.g. ham, bacon, sausage, hot dogs, deli or sandwich and any type of cured and smoked meat) as risk factors for CRC are generally not consistent in the literature. However, many studies, including Johnson et al have reported that excessive consumption of red meat and/or processed meat is associated with an increased risk for CRC [108]. Cancer site thus appears to be a source of variation. In fact, in a meta-analysis, Johnson et al. reported a significant linear dose-response relationship between consuming red meat and developing colon cancer (p = 0.006) [108]. The authors also reported a significant positive relationship between consuming meat and developing the disease (p < 0.001), with a 13% increased CRC risk red meat (RR = 1.13; 95% CI: 1.09 – 1.16) consumers [108]. Recent data suggest that genetic susceptibility may be an effect modifier in the relationship between consuming processed meat and CRC risk [110]. In 2015, after aggregating and reviewing the evidence linking or dissociating the consumption of red and/or processed meat with increased CRC risk, the International Association of Research on Cancer (IARC) classified red meat as "probably carcinogenic to humans" and processed meat as "carcinogenic to humans" [110].

2.2.5.2.2 Cigarette smoking

The association between tobacco cigarette smoking and CRC has been recently established [111]. In November 2009, the IARC concluded that tobacco and cigarette smoking cause CRC [111]. It also appears that, being a current smoker, and/or smoking for longer pack-years increase CRC risk, thus revealing a dose- and time-dependent relationship. Johnson et al. [108] analyzed 7433 CRC cases from 15 studies and reported RR of CRC for smokers (vs non-smokers) at 1.06 (95% CI: 1.03 – 1.08), 1.11 (95% CI: 1.07 – 1.16), 1.21 (95% CI: 1.13 – 1.29) and 1.26 (95% CI: 1.17 – 1.36) for 5, 10, 20 and 30 pack-years respectively. Huxley et al. [109] and Tsoi et al. [112] report a 16% and 20% increased CRC risk among current smokers as compared to non-smokers respectively while Liang et al. reported a 50% increased risk for 60 pack-years [113]. Individuals who smoke cigarettes have an earlier age of onset incidence of CRC [114].

2.2.5.2.3 Physical inactivity and obesity

Physical inactivity and excess body weight are two interrelated lifestyle-associated CRC risk factors that account for about 25% to 33% of all CRC [115]. Higher levels of physical inactivity are generally associated with higher CRC risk [115]. In a random-effects meta-analysis of 21 studies, physically active individuals had a 27% and 26% reduction in the development of both proximal (RR = 0.73; 95% CI: 0.66 – 0.81) and distal tumors (RR = 0.74; 95% CI: 0.68 – 0.80) respectively when compared with least active people [116]. Elsewhere, such an association could not establish for rectal cancers (RR = 0.98; 95% CI: 0.88 – 1.08) [117]. It appears that, engaging in longer hours of physical activity reduces CRC risk among women [118]. In particular, women who engage in more than 21.5 hours per week of physical activity have a 25% lower risk of developing colonic cancer as compared to those who engage in less than 2 hours per week of physical activity (RR = 0.77; 95% CI: 0.58 – 1.01) [118].

Excess body weight (i.e. body mass index – BMI $\geq$ 25kg/m^2) is also associated with increased CRC risk [119]. Obese individuals (BMI $\geq$ 30kg/m^2) have 20% increased risk of developing CRC as compared to persons with normal weight [119]. Previous studies have reported that obesity is associated with 7-60% increased risk for CRC [120]. This association appears stronger among men than among women, and for colonic cancer than for rectal cancer [121]. Obese men have 50% elevated risk of colonic cancer and a 20% elevated risk for rectal cancer while obese women have 20% increased risk colonic cancer and a 10% increased risk for rectal cancer [121]. However, mechanisms that are potentially responsible for the association between excess body weight and CRC risk are incompletely known [121]. In addition, little is known about the interactive effects of obesity and physical activity on CRC risk [122].

2.2.5.2.4 Alcohol consumption

Studies have linked excessive alcohol consumption to increased CRC risk. One such study is the European Prospective Investigation into Cancer and Nutrition (EPIC) study [123]. In this study, 478,732 people were followed for almost a decade, to determine the impact of baseline and lifetime alcohol consumption on the risk of developing CRC. Lifetime alcohol consumption was positively associated with increased CRC risk (hazard ratio, HR = 1.08 for 15 grams/day increase; 95% CI: 1.04 – 1.12), with an elevated risk in the rectum (HR = 1.12; 95% CI: 1.06 – 1.18) than the colon (distal: HR = 1.08; 95% CI: 1.01 – 1.16; proximal: HR = 1.02; 95% CI: 0.92 – 1.12) [123]. Similar results were reported for baseline alcohol consumption [123].

Individuals who have a lifetime average of 2-3 alcoholic drinks daily have a 20% increase in CRC risk as compared to occasional drinkers and non-drinkers, while having more than 3 alcoholic drinks is associated with a 40% higher risk of the disease [124]. In addition, Fedirko et al. analyzed data from 34 case-control and 27 cohort studies and reported RRs of 1.07 (95% CI: 1.04 – 1.10), 1.38 (95% CI: 1.28 – 1.50), 1.82 (95% CI: 1.41 – 2.35) for a 10, 50 and 100 grams/day consumption of alcohol, respectively [125]. A meta-analysis of 14 cohort studies showed that alcohol consumption of 100 gram/week is associated with a 19% increased risk for CRC (RR = 1.19; 95% CI: 1.14 – 1.27) [126]. These associations are generally stronger in men than in women [127]. The reason(s) for such gender differences are still unclear but may be related to hormone-related differences in alcohol metabolism [127].

2.2.5.2.5 Occupation and Occupational Exposures

One of the most suspected risk factors for CRC with inconsistent results in the literature is occupation and occupational exposures. In the last few decades, the role of farming/agricultural occupation in CRC risk has received significant attention [128-129]. In a case-control study, Fredriksson et al reported reduced risks for developing colon cancer among farmers (male: OR = 0.7, 95% CI: 0.4 – 1.0; female: OR = 0.8, 95% CI: 0.5 – 1.2), albeit this was not significant [128]. However, being a gardener was associated increased risk of colon cancer (male: OR = 1.3, 95% CI: 0.5 – 3.5; female: OR = 2.5, 95% CI: 0.8 – 7.9) and this not significant either [128]. Somayeh et al. reported a significant increase in CRC risk among persons with farm-related jobs while controlling for rural-urban residency [adjusted OR = 7.0, 95% CI: 2.19 – 22.38, *p*-value = 0.001] [129]. In a cohort of female farm residents in New York State, Wang et al. reported that persons who have farm jobs and rural residents were associated with significantly lower CRC risk as compared to those with non-farm jobs and urban residents respectively [130]. Carrozza et al. reported that CRC incidence tends to be high in regions with greater farming activity [131].

Occupational exposures have been identified to be associated with CRC risk. Fredriksson et al. reported significant agents for increased CRC risk to include: asbestos (low grade: OR = 2.1, 95% CI: 0.8 – 5.8), organic solvents (low grade: OR = 1.3, 95% CI: 0.8 – 2.0), White spirit (OR = 1.1, 95% CI: 0.5 – 2.3), DDT (OR = 1.3, 95% CI: 0/8 – 2.1), thinners (OR = 1.2, 95% CI: 0.6 – 2.4) and pesticides (OR = 1.6, 95% CI: 0.8 – 3.0), albeit these are not significant [128]. Elsewhere, persons with a history of pesticide exposure had a significantly higher risk of developing CRC (OR = 2.6, 95% CI: 1.1 – 5.9) as well as persons who frequently eat food directly from farms (OR = 4.6, 95% CI: 1.5 – 14.6) [132]. El-Zaemey et al. reported no association between CRC risk and persons exposed to selected dust agents: animal dust (OR = 1.07, 95%

CI: 0.84 – 1.34), Quarts (OR = 0.91, 95% CI: 0.71 – 1.19), and wood dust (OR = 1.08, 95% CI: 0.74 – 1.57) [133].

In conclusion, based on the inconsistent results in the literature regarding farming and its associated occupational exposures and CRC risk, further research is needed to explore this association. In this thesis, I used data from the SRHS to investigate this relationship in rural Saskatchewan where farming is the predominant occupation.

CHAPTER 3 – DATASET DESCRIPTION

In this chapter, I describe the Saskatchewan Rural Health Study (SRHS) data set used in this thesis. The study design and study population as well as the sampling methods used to obtain the study sample are presented in section 3.1 and 3.2 respectively. The tool used for data collection is presented in section 3.3. The theoretical framework as well as the description of the variables used in this thesis are presented in sections 3.4 and 3.5 respectively.

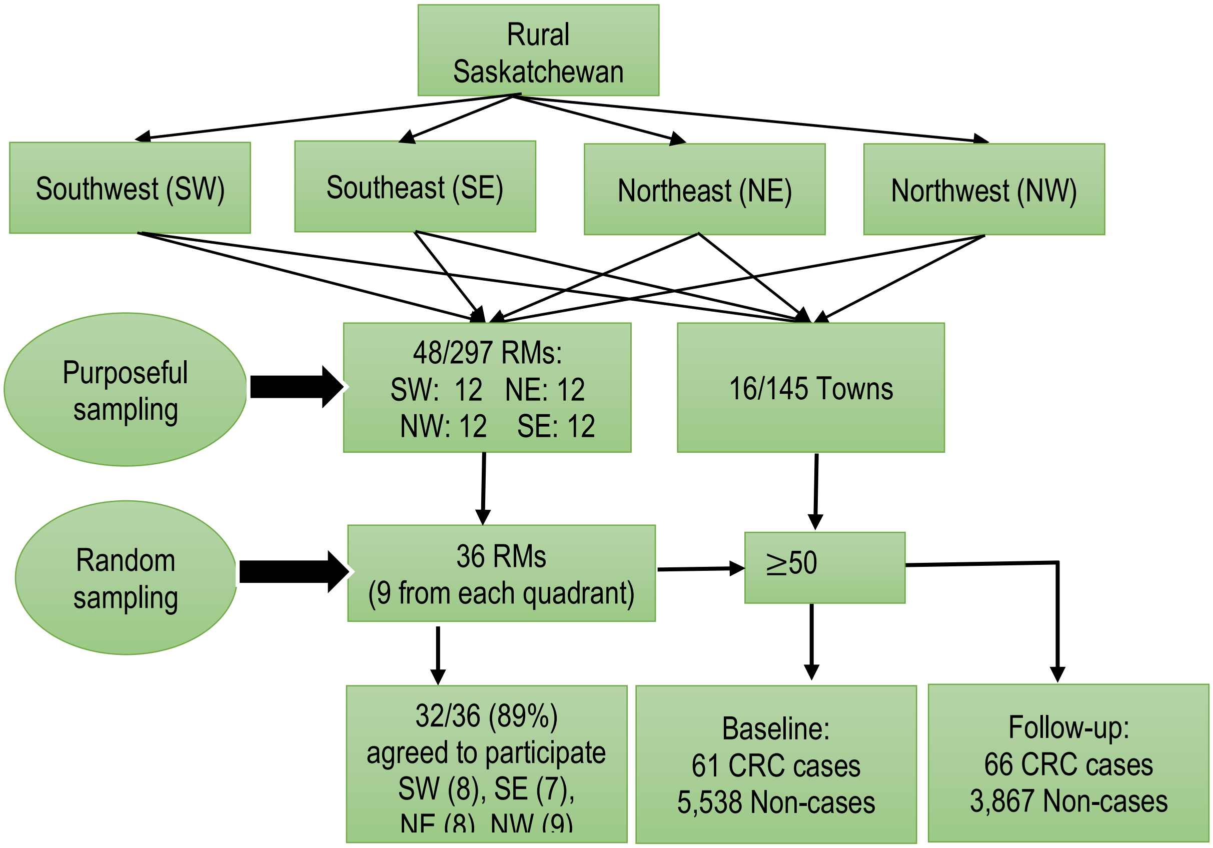
3.1 Study design

The Saskatchewan Rural Health Study (SRHS) is a prospective cohort study of rural dwellers in Saskatchewan, Canada that was conducted in two phases, a baseline survey in 2010 and a four-year follow-up survey in 2014 [134]. The baseline survey was conducted in three stages. In stage one, the southern half of Saskatchewan was partitioned into four quadrants [Southwest (SW), Southeast (SE), Northeast (NE), and Northwest (NE)] using a multistage sampling. Small towns and rural municipalities (RMs) were then selected from these quadrants using Statistics Canada guidelines [135]. A detailed description of the sampling strategy is published elsewhere [134]. In Stage two, a questionnaire was administered to the target population. In the third and final stage, a sub-population of the target population was selected for clinical assessments. Participants of the baseline survey were followed over a four-year period. Data on individual and contextual factors were collected from the baseline survey as well as from the follow-up surveys. This thesis is based on these two datasets. Self-reported doctor-diagnosed CRC cases were identified from the SRHS. The youngest age at diagnosis of the disease was 50 years. The primary group of interest was persons who self-reported doctor-diagnosed CRC and the comparison group was those persons who did not report doctor-diagnosed CRC.

3.2 Study population, Sampling and Recruitment

The SRHS sampled tax-paying households in towns and RMs in Saskatchewan based on a multistage sampling design [134]. The southern half of province was partitioned into quadrants based on Statistics Canada guidelines [135] and 12 RMs were selected in each of the four quadrants. Forty-eight (48) of 297 RMs and 16 of 145 towns located in rural Saskatchewan participated in the SRHS. Thirty-six (36) RMs (9 from each quadrant) were randomly selected. Thirty-two (32) out of the 36 RMs, representing 89% and 15 out of 16 towns, representing 94% consented to participate in the baseline study and provided their mailing addresses to the research team. To maximize the response rate, a modified version of Dillman's method for mail and telephone surveys [136] was used at baseline (RM = 41.9%, small-town = 42.2%) and follow-up (RM = 65%, small-town = 60.2%) surveys. Dillman's method requires that a series of mail correspondence is maintained with all potential participants of the study. Thirty-two RMs [(SW (8), SE (7), NE (8), and NW (9)] and 15 towns [(SW (4), SE (3), NE (2), and NW (6)] in rural Saskatchewan participated in the baseline survey. Baseline questionnaires were mailed to 11,004 households deemed eligible for the study in 2010. Of this number, information was collected on 8,261 individuals aged 18 years or older, nested within 4624 households. In the follow-up survey, information was collected on 4,867 individuals aged 18 years or older, nested in 2797 households via mailed-out questionnaires [137]. Of this number, 4,741 participated in both surveys. There were 126 new individuals who did not take part in baseline survey but participated in the follow-up survey. In the baseline and follow-up surveys, 5,599 and 3,933 individuals were aged 50 years or older respectively. Of these numbers there were 61 CRC cases in the baseline survey and 66 CRC cases (including baseline CRC cases) at follow-up. The present study is based on these CRC cases both at baseline and follow-up for a determination of the prevalence, longitudinal changes in prevalence, and incidence of CRC in rural Saskatchewan. The flow chart in figure 3.1 summarizes the selection of the study sample and sample size.

Figure 3. 1 Flowchart showing selection of the study sample and sample size



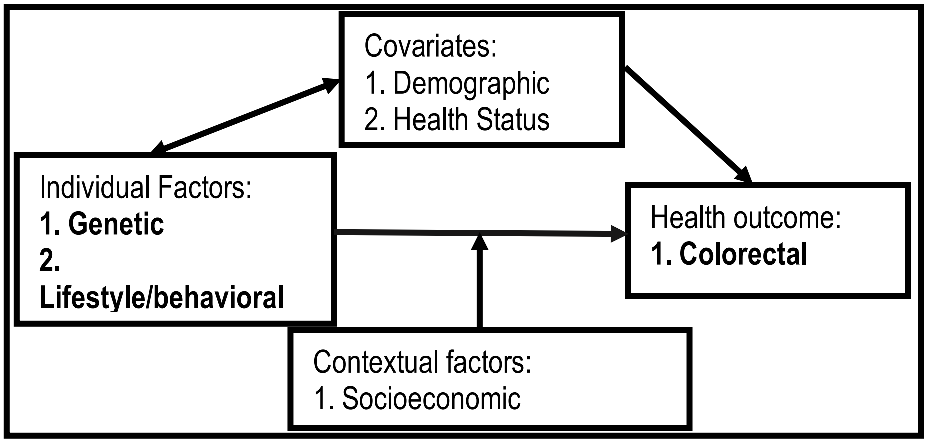
3.3 Data Collection

A household survey questionnaire was developed and used to collect data. Two community members (one from a town and the other from a RM) and the SRHS research team developed the questionnaire. The questionnaire was designed to collect information on individual (e.g. alcohol consumption, cigarette smoking, and obesity), contextual [e.g. socioeconomic, location of home (farm or non-farm), water source, and fuel source], and principal covariates [e.g. demographic variables and body mass index (BMI)]. A pilot survey [134] was conducted to optimize the administration as well as the content of the baseline questionnaire. Based on the pilot survey, some questions were modified in the baseline questionnaire before it was mailed to respondents to be self-administered. A copy of the baseline questionnaire is attached in the appendix section of this thesis (see appendix A).

3.4 Theoretical Framework

To answer the research questions and achieve the study objective stated in section 1.2, we will use Health Canada's "Population Health Framework" (PHF) [138]. The PHF proposes that the interaction of individual and contextual factors may produce different risk levels of an adverse health outcome. Figure 3.2 is a schematic representation of a conceptual framework used to determine how some individual and contextual factors (discussed above) influence CRC risk.

Figure 3. 2 Conceptual framework



Adopted and modified from [139], courtesy of Dr. Will Pickett of Queen's University.

3.5 Study Variables

The variables examined in this thesis were classified into contextual and individual factors as well as covariates. They were obtained directly from the self-administered baseline questionnaire with derivations/reclassification made to some of them where necessary. In the context of this thesis, contextual factors were those exclusive to the rural environment. They are comprised of socioeconomic status and those related to outdoor environmental conditions (e.g. water source, the location of home, etc.). On the other hand, individual factors primarily refer to those factors measuring individual exposures. These included a personal history of smoking and alcohol consumption, family history of cancer, personal history of diabetes, physical activity, and occupational rural exposures related to farming. Covariates included age, BMI, level of education, marital status, gender, and race/ethnic background. The outcome variable, whether or not an individual had ever been diagnosed with CRC by a doctor or primary caregiver (PCG) was determined from the baseline questionnaire.

3.5.1 Operational definitions

3.5.1.1 Contextual factors

Farm: This refers to the location of residences and/or workstations that produce agricultural products intended for sale. The "*Farm*" variable was derived from the responses to the question "*Where is your home located*?" and with responses "*Farm, In-town, Acreage*". For the purposes of this thesis, the farm variable was dichotomized (i.e. "*In-town"* and "*acreage"* were combined to form "*non-farm*") into "*Farm"* and "*Non-farm*" based on question A-1 of the baseline questionnaire (see Appendix A). Categorizing the farm variable into farm and non-farm was necessary because, in rural populations, farming and non-farming exposures are uniquely different from each other.

Household smoking: This was a categorical variable with 2 categories *"Yes/No"* indicating whether or not any household member use tobacco products including cigarette, cigars, and/pipes. It was obtained from the baseline questionnaire based on A-17 question: "*Do any of the people in your house use any of the following tobacco products in your home?*" with response options including "*Cigarette (Yes/No/Don't know)", Pipes (Yes/No/Don't know)*" *and Cigars (Yes/No/Don't know)"* (see Appendix A).

Socioeconomic status: Household income adequacy was used as a proxy measure for socioeconomic status. It was a derived variable with four categories created according to the definition of Statistics Canada [140] based on the total household income and the total number of people in the household based on questions A-2 and A-20 (see Appendix A). Table 3.1 provides a detailed description of the four categories of household income adequacy.

Table 3. 1 Household income adequacy level by income and size of household

Coded value	Income adequacy	Income band	Size of household
1	Lowest income	< $15,000	1 or 2 persons
		< $20,000	3 or 4 persons
		< $30,000	5 or more persons
2	Lower middle income	$15,000 - $29,000	1 or 2 persons
		$20,000 - $39,000	3 or 4 persons
		$30,000 - $59,000	5 or more persons
3	Upper middle Income	30,000 - $59,000	1 or 2 persons
		$40,000 - $79,000	3 or 4 persons
		$60,000 - $79,000	5 or more persons
4	Highest income	≥ $60,000	1 or 2 persons
		≥ $80,000	3 or 4 persons

Quadrant: This refers to the geographical location of the households of participants. It was obtained directly from the baseline questionnaire. This variable was a categorical variable with four levels; SW, SE, NE, and NE that were coded 1 through to 4 respectively. Information on quadrant (i.e. whether RM or small town) was obtained from the 2006 census subdivisions.

Municipality: This refers to whether a household in question was located in a rural municipality (RM) or in a small town in rural Saskatchewan.

Household fuel source - Natural gas and Propane: Information about household fuel source was collected by asking the question "What are the types of fuel sources used to heat your home" *"natural gas (yes/no)"* and "*propane (yes/no)*". This was obtained from the questionnaire based on question A-7 (see Appendix A).

Household water source: This variable had four categories and was a obtained from the questionnaire based on the question "*What is the source of water supply for drinking purposes in your home?*" with options "*Bottled water", "Deep well water (more than 100 feet)", "Shallow well water (less than 100 feet)", Spring, river or creek, Dugout or reservoir, Lake, and Other source*". A new "*water source"* variable was derived and had the categories "*Bottled water, deep well water (more than 100ft), shallow well water (less* *than 100ft)*" and all the remaining categories were combined into "*Other source*". This was based on question A-37 of the baseline questionnaire (see Appendix A).

Mildew odor or musty smell: Information on the household indoor environment was collected using the question "*Does your home (including basement) frequently have mildew or musty smell?*" with answers "*Yes/No*". This was obtained from the questionnaire based on question A-14 (see Appendix A).

3.5.1.2 Individual factors

Personal history of smoking: Information about the individual history of smoking was collected. The smoking variable was a dichotomous variable and was self-reported. This was obtained from the questionnaire based on the question "*Have you ever smoked cigarettes?*" with responses "*Yes/No*" using question B-36. For the purpose of this thesis, the smoking variable was further categorized into 3 categories; current smoker, ex-smokers, and never-smoker. This was a derived variable and was obtained using questions B-37 through to B-45 (see Appendix A).

Alcohol consumption: Data about the history and frequency of alcohol consumption was collected by asking the question "*During the past 12 months, how often did you drink alcoholic beverages?*". For the purpose of analysis in this study, a new *alcohol* variable was derived that categorized responses to this question into 3 categories which were "*Never*", "*Everyday*" and all the remaining categories were combined to obtain the "*Occasionally*" category based on question B-46 (see Appendix A).

Personal and Family history of cancer: Information about the personal as well as family history of cancer was also obtained. Personal history of cancer was obtained directly from the baseline questionnaire by asking the question "*Has a doctor or primary caregiver ever said you have cancer?*" with response options "*Yes/No/Don't know*". This was obtained from question B-50 (see Appendix A). On the other hand, family history of cancer was also obtained by asking "*Have the following members of your biological family ever had cancer?*" and the list included "*Father*, *Mother* and *Brother/Sister*" *with response options (Yes/No/Don't know)".* These were obtained directly from the baseline questionnaire using question B-56 (see Appendix A).

For the purpose of this thesis, a new variable was derived to recategorize the *family history of cancer* variable into four categories which included "*Father cancer*", "*Mother cancer*", "*Both parents cancer*", as well as "*Sibling cancer*".

Diabetes: Information about the personal history of diabetes was also collected. Personal history of diabetes was obtained directly from the baseline questionnaire by asking the question "*Has a doctor or primary caregiver ever said you have diabetes?*" with response options "*Yes/No/Don't know*". This was obtained from question B-50 (see Appendix A).

***Physical activity*:** Personal of physical activity was collected by asking the question "*Do you exercise?*" with a dichotomous response option "*Yes/No*". This was obtained from the baseline questionnaire using question B-27 (see Appendix A).

***Early life exposure to farm*:** Information about early life exposure to a farming environment was obtained from the baseline questionnaire. To ascertain whether a respondent ever lived on a farm, the question "*Have you ever lived on a farm?*" with response options "*Yes/No*" was asked. On the other hand, "Did you live on a farm during your first year of life" was used to ascertain whether a respondent was exposed to a farming environment in their first year of life ("*Yes*") or not ("*No*"). These were obtained using questions B-31 and B-32 (see Appendix A).

Occupational rural exposure: Information about whether respondents were exposed to selected occupational exposures was collected by asking the question "*Have you been exposed to any of the following in your workplace?*". The list of exposures was; livestock, stubble smoke, fungicides, herbicides, molds, oil/well fumes, radiation, solvent fumes, welding fumes, grain dust, wood dust, diesel fumes, asbestos dust, insecticides, mine dust, and other_specify. These were obtained from the baseline questionnaire based on question B-58 (see Appendix).

3.5.1.2 Covariates

Age**:** The SRHS collected data on individuals age 18 years or older at baseline. For the purposes of this thesis, individuals aged 50 years or older were extracted from the SRHS dataset. We chose this age cut-off point because that was the youngest age of diagnosis for the CRC patients. Information about age was obtained directly from the baseline questionnaire and the age variable was continuous. Age was categorized into four categories which included 50-59 years (reference category), 60-69 years, 70-79 years and 80 years or older.

Body Mass Index (BMI): The *BMI* was derived from the baseline questionnaire. It was calculated using the formula:

$$\left[\frac{Weight\ in\ kilograms}{(Height\ in\ centimeters)^2}\right] \mathrm{X}100{,}000$$

Respondents self-reported their and weight in the baseline questionnaire. Participants were asked "*What is your height? _kg or ft and in*" and "*What is your weight?* _kg or lbs". For the purpose of data analysis in this thesis, the baseline BMI was recorded into three (3) categories which included: normal weight (reference category) (0 - < 25kg/m^2), overweight (25 - 30 kg/m^2), and obese (more than 30 kg/m^2). These were obtained from the baseline questionnaire using question B-6 and B-7 (see Appendix A).

Educational status**:** The level of education was self-reported by study participants and was obtained from the baseline questionnaire using the question "*What is your highest level of education?*" with response options "< *high school", "Completed high school", "Completed university",* and *"Completed post-secondary education other than above*". For the purpose of the present study, a new *education* variable was derived with two categories "*Grade 12 or below*" and "*Beyond Grade 12*". The "< *high school*" and "*Completed high school*" were combined into "Grade 12 or below" while the "*Completed university*" and "*Completed post-secondary education other than above*" categories were recoded into "Beyond Grade 12". These were obtained from the baseline questionnaire using question B-4 (see Appendix A).

Marital status: Respondents self-reported their status by answering the question "*What is your marital status?*" based on question B-8 (see Appendix A).

Gender: This was a dichotomous variable with categories "*Male*" (reference category) and "*Female*" using on question B-3 (see Appendix A).

Race/Ethnic background: Information about the race/ethnic background was asked all respondents using the question "What is your ethnic background?" and the response options included *"Caucasian", "First Nation", "Metis", and "Other_please specify"*. The "*Other_please specify*" category included all those who did not belong to any of the other categories and those who also did not know their ethnic background. Due to small frequencies in the last three categories, the *race* variable was recoded into "*Caucasian*" or "*non-Caucasian*". "*First Nation, Metis, Other_please specify"* categories were combined to produce the "*non-Caucasian*". This achieved using question B-5 (see Appendix A).

3.5.1.3 Outcome

The primary outcome variable in this thesis was self-reported doctor/PCG-diagnosed colorectal cancer (CRC). This was obtained from the baseline questionnaire using the question "*Has a doctor or primary caregiver ever said you have cancer? If yes, please specify cancer type*". For the purposes of analysis, the "*colorectal cancer*" variable was recoded into a dichotomous variable ("*Yes/No*"), such that CRC cases belong to the "*Yes*" category while CRC non-cases belong to the "*No*" category. This was achieved based on question B-50 (see Appendix A).

CHAPTER 4 – METHODS

In longitudinal data analysis, appropriate methods must account for the within-subject correlations that may exist due to the presence of more than one observation per subject. In this chapter, models accounting for this within-subject correlations as used in this thesis are discussed. In particular, models for discrete longitudinal data analysis are presented and for each objective (see section 1.2 above), appropriate models are discussed. Marginal modeling for discrete longitudinal data was used to address objective 1 in order to determine significant risk factors for the prevalence of CRC. Cox's PH and discrete PH models were used to address objective 2 in order to determine significant risk factors for the incidence of colorectal cancer.

The remainder of this chapter is organized into Part I and II. Part I is organized as follows: In section 4.1, models for longitudinal data analysis are provided and the univariate quasi-likelihood function is described in section 4.2. Generalized linear models (GLMs) to handle repeated measurements in longitudinal data are provided in section 4.3 with a particular focus of the multivariate quasi-likelihood function. We extend the GLMs in section 4.4 with a focus on marginal modeling for longitudinal data. Statistical application of these methods to determine significant factors for CRC prevalence is provided in section 4.5. In Part II, a discussion of models for survival analysis is provided in section 4.6.

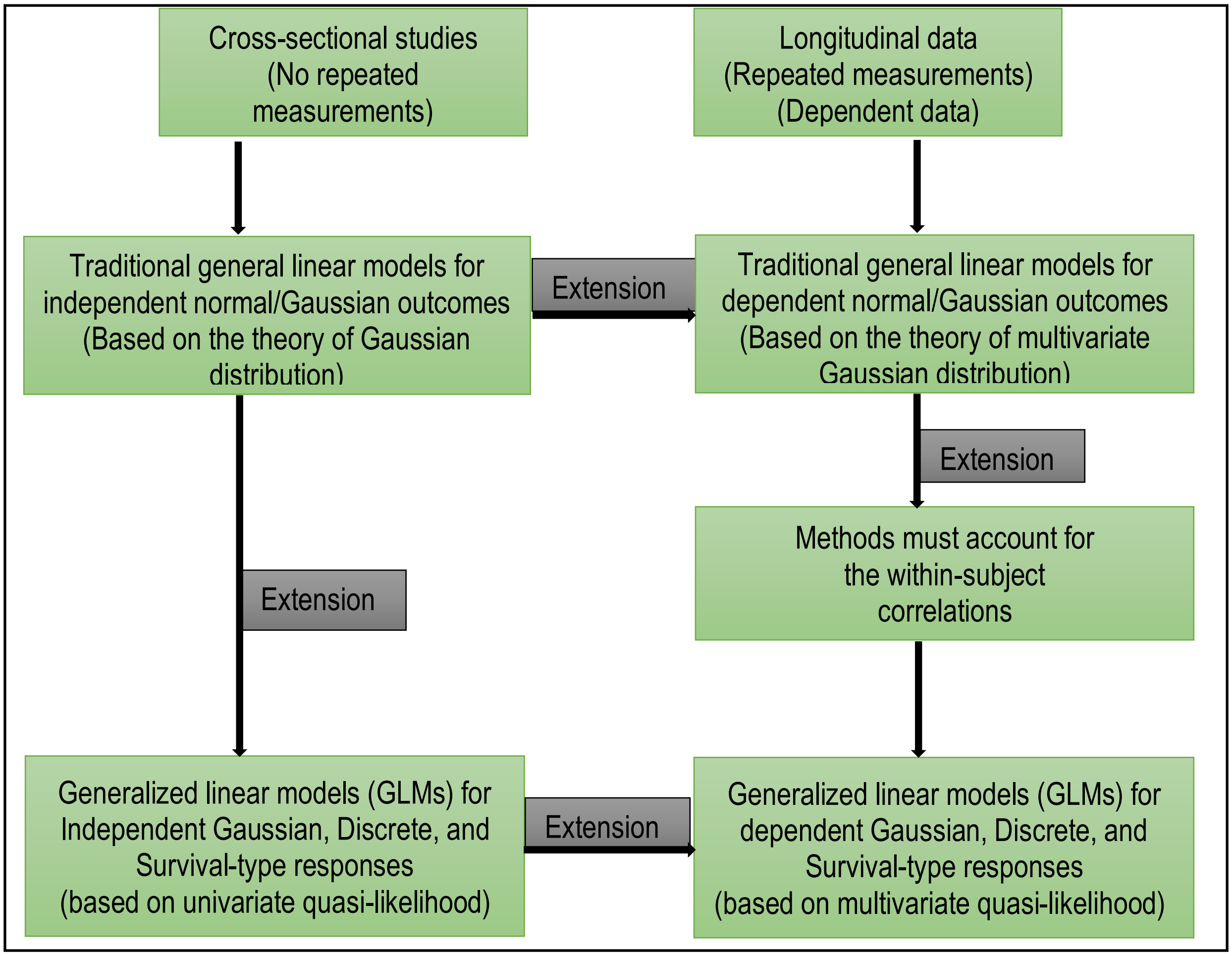
PART I

OUTCOME VARIABLE: DISCRETE/DICHOTOMOUS

4.1 Models for Longitudinal Data Analysis

In most medical and epidemiological studies, it is usually of interest to determine the relationship between a response/outcome variable Y, and an explanatory or a set of explanatory variables X. In cross-sectional studies, classical/general linear models using the mathematical theory of normal distribution are developed to model independent and Gaussian responses [2]. However, to model independent responses for both Gaussian and non-Gaussian data, a specialized class of regression models called generalized linear models (GLMs) are used [43]. GLMs are based on the theory of quasi-likelihood and are used to extend the scope of determining the relationship between responses and explanatory variables [43]. This shows that general/classical linear models for independent data can be extended using GLMs for independent responses. Classical regression methods are not suitable for analyzing data from longitudinal studies because they do not account for the within-subject correlation that exists in longitudinal data [3]. In longitudinal studies, repeated measurements on the same subject tend to be correlated with each other and as a result, the independence assumption in GLMs is violated. In such instances, appropriate models are developed to take into account the inter-dependence among repeated measurements in order to make an accurate determination of the relationship between Y and X and any inferences thereof. Models based on the mathematical theory of multivariate Gaussian distribution are developed as an extension to classical linear models for longitudinal data [2]. On the other hand, models based on the mathematical theory of multivariate quasi-likelihood are developed as an extension to GLMs which are based on univariate quasi-likelihood [2]. For example, the generalized estimating equations (see section 2.1.2.1) are based on the multivariate quasi-likelihood and can be used to model both Gaussian and non-Gaussian outcomes [3]. Figure 4.1 shows a schematic representation of statistical methods commonly used for independent and dependent outcomes in longitudinal data analysis.

Figure 4. 1 Statistical methods available for independent and dependent outcomes in longitudinal data analysis (adopted and modified from [2] with authors permission)


Cross-sectional studies
(No repeated
measurements)
Longitudinal data
(Repeated measurements)
(Dependent data)
Traditional general linear models for
independent normal/Gaussian outcomes
(Based on the theory of Gaussian
distribution)
Extension
Traditional general linear models for
dependent normal/Gaussian outcomes
(Based on the theory of multivariate
Gaussian distribution)
Extension
Methods must account for
the within-subject
correlations
Extension
Generalized linear models (GLMs) for
Independent Gaussian, Discrete, and
Survival-type responses
(based on univariate quasi-likelihood)
Extension
Generalized linear models (GLMs) for
dependent Gaussian, Discrete, and
Survival-type responses
(based on multivariate quasi-likelihood)

4.2 The Univariate Quasi-Likelihood Function

Suppose we have a vector of responses,$\mathbf{Y}$ $(y_i = 1,2, \dots n)$ which are independent with expectation μ_i and variance $V(\mu_i)$ where V is made up of some known functions. Also, suppose that μ_i is a function of the regression coefficients $\boldsymbol{\beta}$ (i.e. $\boldsymbol{\beta} = \beta_{1,}\beta_{2,\dots}\beta_k, \dots \beta_p$). Then for a single component of $\mathbf{Y}$, Wedderburn [42] defined the quasi-likelihood function $K(y_i, \mu_i)$ by the relation:

$$\frac{\partial K(y_i, \mu_i)}{\partial \mu_i} = \frac{y_i - \mu_i}{V(\mu_i)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . \ldots\ldots\ldots\ldots . (4.1)$$

or equivalently,

$$K(y_i, \mu_i) = \int^{\mu_i} \frac{y_i - \mu'_i}{V(\mu'_i)} d\mu'_i + \text{some function of } y_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (4.2)$$

The univariate quasi-likelihood function $K(y_i, \mu_i)$ has four properties which are similar to those of the log likelihood function (i.e. the score). These properties include;

(i) $E\left(\frac{\partial K}{\partial \mu}\right) = 0$

(ii) $E\left(\frac{\partial K}{\partial \beta_i}\right) = 0$

(iii) $E\left(\frac{\partial K}{\partial \mu}\right)^2 = -E\left(\frac{\partial^2 K}{\partial \mu^2}\right) = \frac{1}{V(\mu)}$, *and*

(iv) $E\left(\frac{\partial K}{\partial \beta_i}\frac{\partial K}{\partial \beta_j}\right) = -E\left(\frac{\partial^2 K}{\partial \beta_i \partial \beta_j}\right) = \frac{1}{V(\mu)}\frac{\partial \mu}{\partial \beta_i}\frac{\partial \mu}{\partial \beta_j}$

Details of the proofs to these properties and other relevant information about the quasi-likelihood function are provided by Wedderburn [57]. To obtain the quasi-likelihood estimating equation for $\boldsymbol{\beta}$, the regression coefficients, we differentiate $K(y_i, \mu_i)$. The resulting estimating equation is called the quasi-score function and can be written as;

$$\boldsymbol{U}(\beta) = \boldsymbol{D}'\boldsymbol{V}^{-1}(\boldsymbol{Y} - \mu) = 0 \ldots\ldots\ldots\ldots\ldots.\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (4.3)$$

where D is an $n \text{ x } p$ matrix with the $(i,j)^{th}$ element as $\frac{\partial \mu_i}{\partial \beta_j}$ and $V = \text{Var}(Y_i)$ is an $n \text{ x } n$ diagonal matrix with $\text{Var}(\mu_i)$ as its i^{th} diagonal entry. An alternative parameterization of equation (4.3) is given as;

$$\boldsymbol{U}(\beta) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \beta_k}\right) [\text{Var}(Y_i)^{-1}]\,(Y_i - \mu_i) = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (4.4)$$

To account for the within-subject correlations, generalized linear models (GLMs) (based on the mathematical theory of quasi-likelihood) are ‘generalized’ or extended to take into account repeated measurements for each subject. Currently, the generalized estimating equations (GEE) (the multivariate analog of the quasi-likelihood) is used to account for the within-subject correlations arising due to the longitudinal design.

4.3 Generalized Linear Models (GLMs) for Longitudinal Data Analysis

Generalized Linear Models (GLMs) have been extended to handle repeated measurements in longitudinal data. Depending on how to account for the within-subject correlations, there are different regression models for longitudinal data analysis [2]. Three common extended models of the GLMs for longitudinal data analysis include; (i) marginal/population-average (P-A), (ii) transition/response conditional, and (iii) random effects/subject specific models [2-3]. For the purpose of this thesis, I focused on marginal models because my main interest is to model the mean CRC outcome of the population as a function of only the explanatory variables and covariates and not of any previous response or random effects. Because of this, I require only a regression model for the mean response (i.e. CRC outcome) and not the full distributional assumptions associated with the vector of repeated responses [2]. GEE is the multivariate analog of the univariate quasi-likelihood theory. The GEE approach is used to analyze a discrete CRC outcome (yes/no) in this thesis. A detailed formulation of the GEE methodology is reviewed next.

4.3.1 The Generalized Estimating Equations (GEE) Methodology

In this section, I present the formulation of the GEE methodology. The notations and definitions of Liang and Zeger [3] will be adopted unless otherwise stated. To understand the mathematical theory presented in this section, the definitions for the notations of vectors and matrices are presented in Appendix C of this thesis.

Let $\boldsymbol{Y}_{it} = (y_{i1}, y_{i2}, \dots, y_{in_i})^{\mathrm{T}}$ be a $n_i\mathrm{x}1$ vector of responses and $\boldsymbol{X}_i = (x_{i1}, x_{i2}, \dots, x_{in_i})^{\mathrm{T}}$, the $n_i\mathrm{x}p$ matrix of covariate for the i^{th} subject ($i = 1,2 \dots, K$), then the marginal density of y_{it} is given as;

$$f(y_{it}) = \exp[\{y_{it}\theta_{it} - a(\theta_{it}) + b(y_{it})\}\phi] \dots \dots \dots \dots \dots \dots \dots \dots . . \dots \dots \dots \dots \dots (4.5)$$

where $\theta_{it} = h(\eta_{it})$, and $\eta_{it} = x_{it}\beta$. The mean and the variance of y_{it} are given by;

$$\left.\begin{aligned} E(y_{it}) &= a'(\theta_{it}) \\ \mathrm{var}\,(y_{it}) &= {a''(\theta_{it})}\big/{\phi} \end{aligned}\right\} \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots . (4.6)$$

Liang and Zeger formulated the ‘independence’ estimating equations based on a working assumption that, repeated observations for an experimental unit are independent of each other, which led to the estimation of regression parameters and their covariance matrices [4]. Under this “independence” working assumption, Liang and Zeger proposed that the score equations associated with the quasi-likelihood function are of the form

$$U_I(\beta) = \sum_{i=1}^{K} X_i{}^T \Delta_i S_i = 0 \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots \dots .. (4.7)$$

where Δ_i=diag$(d\theta_{it}/d\eta_{it})$ is an $n\text{x}n$ matrix while $S_i = Y_i - a'(\theta_{it})$ is an $n\text{x}1$ matrix for the i^{th} experimental unit [4]. The solution to this equation (i.e. Eq. 4.7) yields consistent estimates $\hat{\beta}_I$ and $var(\hat{\beta}_I)$ of β and $var(\beta)$ respectively only when the regression model for $E(y)$ is correctly specified. Moreover, missing data should be missing completely at random (MCAR) for this ‘consistency’ to hold [4]. A primary drawback of $\hat{\beta}_I$ is that it may not have high efficiency in instances where the autocorrelation is huge [4]. To overcome these limitations, Liang and Zeger extended the ‘independence’ estimating equations to the generalized estimation estimating equations (GEE) [4]. The GEE methodology accounts for the within-subject correlations, thus resulting in consistent estimators with higher efficiency as well [4]. To illustrate this, let $R(\alpha)$ be an $n\text{x}n$ “working” symmetric correlation matrix, where $\boldsymbol{\alpha}$ is an $s\text{x}1$ vector fully characterizing $\boldsymbol{R}(\boldsymbol{\alpha})$. If $\boldsymbol{R}(\boldsymbol{\alpha})$ is actually the true correlation matrix for the responses (i.e. $\boldsymbol{Y}_i$’s), then $\text{cov}(\boldsymbol{Y_i})$ =V_i, where V_i is defined as;

$$V_i = \frac{A_i{}^{\frac{1}{2}} R(\alpha) A_i{}^{\frac{1}{2}}}{\phi} \dots \dots \dots \dots \dots \dots \dots \dots .. \dots \dots \dots \dots \dots \dots (4.8)$$

where A_i = diag$\{a''(\theta_{it})\}$ and ϕ is a scale parameter. As a result, Liang and Zeger [4] defined the general estimating equation as;

$$\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} S_i = 0, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.9)$$

where $$D_i = {d\{a'_i(\theta)\}}\Big/{d\beta} = A_i \Delta_i X_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.10)$$

A_i is a diagonal matrix with $a''(\theta_{it})$ as elements along the main diagonal.

It is worthy of note that, when $R(\alpha)$ is specified as the identity matrix, the estimating equation specified in $(\mathrm{Eq.}\,4.9)$ easily reduce to the “independence” estimating equation.

For a univariate GLM, the estimating equation for quasi-likelihood is of the form;

$$\sum_{i=1}^{K} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)^{\mathrm{T}} = \boldsymbol{V}^{-1}\{\boldsymbol{\mu_i}(\boldsymbol{\beta}), \widetilde{\boldsymbol{\alpha}}\}[y_i - \boldsymbol{\mu_i}(\boldsymbol{\beta})] = 0 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.11)$$

where $\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ and $\tilde{\alpha} = (\alpha', \phi)'$ and ϕ is an unknown dispersion parameter.

Liang and Zeger developed the multivariate generalized equation analog of equation (4.11) as;

$$\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{\mathrm{T}} \, [y_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})] = 0, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.12)$$

where D_i is defined in equation (4.10).

Liang and Zeger [3], as the number of clusters increases infinitely, the regression estimator derived through the GEE approach, denoted $\boldsymbol{\beta}_{GEE}$, is asymptotically multivariate Gaussian with mean zero and covariance matrix ($\boldsymbol{\Sigma}$), defined by $\boldsymbol{V}_{GEE}$ [i.e. $\boldsymbol{\beta}_{GEE} \sim N(0,\ \boldsymbol{V}_{GEE})$]. $\boldsymbol{V}_{GEE}$ is defined as;

$$\boldsymbol{V}_{GEE} = \lim_{K\to\infty} K\left(\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} D_i\right)^{-1} \left\{\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} \mathrm{cov}(Y_i)\, V_i^{-1} D_i\right\} \left(\sum_{i=1}^{K} D_i^{\mathrm{T}} V_i^{-1} D_i\right)^{-1} \ldots\ldots (4.13)$$

The variance of $\boldsymbol{\beta}_{GEE}$, (denoted $\boldsymbol{V}_{GEE}$) is easily obtained by replacing $\mathrm{cov}(Y_i)$ by $S_i S_i^{\mathrm{T}}$ as well as replacing $\boldsymbol{\beta}$, ϕ, and α by their respective estimates in equation (4.13) above. For $\widehat{\boldsymbol{\beta}}_{GEE}$ and $\widehat{\boldsymbol{V}}_{GEE}$ to be consistent, one must correctly specify only the mean since the choice of R is inconsequential. Zeger and Liang [4] presented a review of statistical methods for analyzing longitudinal data analysis using discrete and continuous outcomes. The GEE approach discussed above is based on the multivariate quasi-likelihood function which is presented next.

4.3.2 The Multivariate Quasi-Likelihood Function

The multivariate quasi-likelihood function as proposed by Liang and Zeger [3] is a generalization of the univariate quasi-likelihood proposed by Wedderburn [42]. Let $\boldsymbol{\Sigma}_i$ $(i = 1,2,\dots,m)$ be an $n_i \mathrm{x} n_i$ vector of responses Y_i, associated with R_i, an $n_i x n_i$ correlation matrix. If $V_j{}^{1/2}(j = 1,2,\dots,n_i)$ is a diagonal matrix whose j^{th} diagonal entry is $\sqrt{[\phi V(\mu_{ij})}$ such that $Var(Y_{ij}) = \phi V(\mu_{ij})$ and $Var(Y_i) = \boldsymbol{\Sigma_i} = V_i{}^{1/2}\mathrm{R}_i V_i{}^{1/2}$, then the multivariate quasi-likelihood function is given as;

$$\boldsymbol{U}(\beta) = \sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta_k}\right) [\boldsymbol{\Sigma}_i(\alpha)]^{-1} (Y_i - \mu_i) = 0 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.14)$$

where $\frac{\partial \mu_i}{\partial \beta} = [(\partial \mu_{ij}/\partial \beta_k)]$ $(k = 1,2,\dots,p)$ is a $p\mathrm{x}n_i$ matrix, $\boldsymbol{\Sigma}_i(\alpha) = V_i{}^{1/2}\mathrm{R}_i V_i{}^{1/2}$ is an $n_i \mathrm{x} n_i$ matrix, and $Y_i - \mu_i$ is an $n_i \mathrm{x}1$ vector. Over dispersion is accounted for and is estimated by the dispersion parameter ϕ using $\hat{\phi} = \frac{1}{N-p}\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{y_{ij}-\mu_{ij}}{\sqrt{[Var(\mu_{ij})}}\right]$ where N is the total number of observations is and p is the number of regression coefficients. Also, $E(y_{ij}) = \mu_{ij}$ and $g(\mu_{ij}) = \eta_{ij} = \boldsymbol{X'}_{ij}\boldsymbol{\beta}$ where $X_{ij} = [(X_{ij1}, X_{ij2}, \dots, X_{ijp})]'$, and $g(\cdot)$ is the link function. Given these, the multivariate quasi-likelihood estimating equations or GEEs are given as;

$$\sum_{i=1}^{m} \left(\frac{\partial \mu_i}{\partial \beta_j}\right) [\boldsymbol{\Sigma}_i(\alpha)]^{-1} (Y_i - \mu_i) = 0. \,.\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots (4.15)$$

Equation (4.15) is the multivariate generalization/extension of the univariate quasi-likelihood function in equation (4.2) above.

4.4 Marginal models for Longitudinal Data Analysis

When inferences about population-averages are of interest, marginal models are the most appropriate [38, 187]. The term “marginal” is used to highlight the fact that the mean or population-average response modeled is a function of only the covariates and does not depend on other responses or random effects. Marginal models provide group-level (as against individual-level) information and the regression coefficients are interpreted for the group rather than the individual coefficients [187]. In marginal modeling, the regression model for the mean and the within-subject association (e.g. covariance) are modeled separately [187].

Let Y_{ij} be a continuous, binary, or a count response variable for the i^{th} subject at the j^{th} time denoted t_{ij}. For n_i repeated measurements obtained on the i^{th} subject, there exist a $p\text{x}1$ vector of covariates X_{ij}, associated with each Y_{ij}. Given these, the marginal model has the following three-part specification [187]:

1. The marginal expectation of Y_{ij}, denoted $E(Y_{ij}) = \mu_{ij}$, depends on only the covariates through a specified link function given by;

$$g(\mu_{ij}) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p X_{pij} = \boldsymbol{X'}_{ij}\boldsymbol{\beta} \ldots\ldots\ldots\ldots\ldots..(4.16)$$

   where $g(\cdot)$ is the link function (e.g. identity for Gaussian responses, logit for binary outcomes, and log for counts).

2. The marginal variance of each Y_{ij}, given the covariates, depends on the marginal mean through

$$Var(Y_{ij}|\boldsymbol{X}_{ij}) = \phi v(\mu_{ij}) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(4.17)$$

   where $v(\mu_{ij})$ is a known variance function and ϕ is a dispersion parameter which explains the variation in the Y_{ij}’s that is not explained by $v(\mu_{ij})$. ϕ may be known or estimated.

3. The within-subject association or covariance among the vector of repeated measurements is a function of the marginal means and also of an additional association parameter α (that may also need to be estimated) as;

$$Cov(Y_{ij}, Y_{ik}) = \rho(\mu_{ij}, \mu_{ik}, \alpha) \quad (j < k\ ), \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots. (4.18)$$

where ρ is some known function. In particular, when α represents the pairwise correlations among the Y_{ij}'s, the covariances among the Y_{ij}'s depend on ϕ, $\mu_{ij}(\beta)$, and α through;

$$Cov(Y_{ij}, Y_{ik}) = s.d\ (Y_{ij})Corr(Y_{ij}, Y_{ik})s.d(Y_{ik})$$

$$= \left[\sqrt{\phi v(\mu_{ij})}\right][Corr(Y_{ij}, Y_{ik})][\sqrt{\phi v(\mu_{ik})}]$$

where $s.d\ (Y_{ij})$ represents the standard deviation of Y_{ij}.

4.4.1 Marginal Models for Discrete Outcomes

As mentioned in section 4.4, marginal models allow for the modeling of the marginal mean or expectation of the Y_{ij}'s, $E(Y_{ij})$ as a function of only the covariates and not other responses or random effects. Traditional logistic regression models for dichotomous and polytomous outcomes and Poisson regression for counts/rates are two commonly used models for analyzing discrete outcomes. In this thesis, marginal models were fitted to a dichotomous CRC outcome.

Let Y_{ij} a binary outcome variable (e.g. presence or absence of *CRC*) for the i^{th} subject ($i = 1,2, \ldots, m$) corresponding to the j^{th} time point ($j = 1,2, \ldots, n_i$). In addition, let $\boldsymbol{X}_{ij}$ be a $1xp$ matrix of covariates such that the first element is 1, thus representing the intercept. If $E(Y_{ij}) = \mu_{ij} = \mathcal{P}r(Y_{ij} = 1)$, then the marginal or P-A model for the binary Y_{ij} is a logistic regression model of the form;

$$logit[E(Y_{ij})|\boldsymbol{X}_{ij}] = logit(\mu_{ij}) = logit[\mathcal{P}r(Y_{ij} = 1)|\boldsymbol{X}_{ij}\ ] = \boldsymbol{X'}_{ij}\boldsymbol{\beta}. \ldots\ldots\ldots\ldots\ldots (4.19)$$

The regression coefficients in the marginal model specified in equation (4.19) have population-averaged interpretations where the averaging is over all subjects within the different subgroups of the population [187]. Each element of $\boldsymbol{\beta}$ represents the change in the log-odds of developing a specified outcome resulting from a per unit change in the particular covariate, for all the other known subgroups and fixed covariate values. In general, the interpretation of any of the components of $\boldsymbol{\beta}$, say β_k is in terms of the changes in the population-averaged response or the marginal expectation of the Y_{ij}'s for a unit change in the associated covariate [187]. In this case, e^{β_k} measures the prevalence odds ratio (POR) which is the ratio of the odds of developing the outcome (e.g. disease) among persons who have a specified condition to the odds of developing the outcome among persons who do not have the condition [28]. Mathematically, we write the POR as;

$$e^{\beta_k} = \frac{\mathcal{P}r(Y_{ij}=1|X_{ijk}=1)/\mathcal{P}r(Y_{ij}=0|X_{ijk}=1)}{\mathcal{P}r(Y_{ij}=1|X_{ijk}=0)/\mathcal{P}r(Y_{ij}=0|X_{ijk}=0)} \equiv \frac{\mathcal{P}r(Y_{ij}=1|X_{ijk}=1)\mathcal{P}r(Y_{ij}=0|X_{ijk}=0)}{\mathcal{P}r(Y_{ij}=1|X_{ijk}=0)\mathcal{P}r(Y_{ij}=0|X_{ijk}=1)}.\dots\dots\dots.(4.20)$$

Worthy of note is the fact that, the within-subject association (e.g. correlation) between the Y_{ij}'s does not alter the interpretation of $\boldsymbol{\beta}$ [for instance, a compound symmetry correlation for count data i.e. $Corr(Y_{ij}, Y_{ik}) = \alpha$].

The details of the GEE approach for marginal modeling are provided by Liang and Zeger [3] and Diggle [20].

4.5 Research Question One (1): Marginal Modeling Approach

The main focus of the first objective in this thesis was to determine the crude and adjusted prevalence of self-reported doctor-diagnosed CRC and associated risk factors among farm and non-farm residents of rural Saskatchewan using the SRHS data set. The method used to calculate the crude prevalence rates is presented below.

4.5.1 Statistical Application: Research Question One (1)

4.5.1.1 Estimating Crude Prevalence of CRC

Prevalence proportion is defined to be “the proportion of people in a population that has a disease” [141]. Mathematically, prevalence proportion can be written as;

$$\frac{Number\ of\ people\ who\ have\ the\ disease\ in\ the\ population\ (x)}{Population\ size\ (N)} \text{x } 100 \ldots\ldots\ldots\ldots\ldots\ldots..(*)$$

The above formula was used to calculate crude prevalence rates of CRC for farm and non-farm rural residents each of the four quadrants of rural Saskatchewan, both at baseline and follow-up.

4.5.1.2 Estimating adjusted Prevalence of CRC

The adjusted prevalence of CRC was estimated by fitting a dichotomous regression model adjusted for covariates. Marginal logistic regression models using the generalized estimating equations (GEE) approach were fitted using the SAS procedure GENMOD to identify significant risk factors for colorectal cancer (CRC) prevalence both at baseline and follow-up. The GEE approach is based on the mathematical theory of multivariate quasi-likelihood. The clustering effects arising from having more than one subject per household was accounted for using GEE, assuming an exchangeable correlation matrix. This was achieved by specifying TYPE=EXCH option in the REPEATED statement. The exchangeable correlation structure assumes that the correlation between two (2) repeated outcome measurements are the same and entries of the principal diagonal of the covariance matrix are the same.

I also determined longitudinal changes in the prevalence of CRC by fitting marginal logistic regression models to account for two layers of complexity: first, the effects of having more than one individual in a household; and second; the repeated measures over time. To do that, a new SAS code was written using the GENMOD procedure to simultaneously account for these two layers of complexity by nesting the within-subject effect in the repeated measurements variable and specifying that in the REPEATED SUBJECT statement. Prior to that, baseline and follow-up data were combined and restructured into a long format. A dichotomous index variable was then created with two categories 0 and 1 representing baseline and follow-up respectively. This index variable was specified in the within subject statement.

A univariable logistic regression model used to examine the association between the location of residence (farm or non-farm) and CRC risk. A multivariable logistic regression model was used to assess the relationship between a binary CRC outcome (presence/absence) and a set of predictor variables. Multilevel logistic regression models based on GEE [individual (1st level) nested within households (2nd level)] was also used to assess the relationship between individual and contextual factors. Standard model building strategies were employed to select the final model. In the univariable analysis, variables with p-values <0.25 were candidates for the multivariable analysis. Overall, statistical significance was determined at 5% and only variables with p-values<0.05 were included in the final multivariable model. Important covariates such as gender and BMI were retained in the multivariable model even if they were not statistically significant in the univariable analysis stage. Testing for potential interactions and confounders was conducted in the model building process. The QIC (quasi-likelihood under the independence model criterion) was used to select the most parsimonious model in this thesis. Odds ratios (ORs) and their corresponding 95% confidence intervals (CI) were used to describe the strength of associations.

PART II

OUTCOME VARIABLE: TIME-TO-EVENT/SURVIVAL TIME

4.6 Models for Survival Data Analysis

4.6.1 Introduction

Survival data frequently arise in epidemiological studies and clinical trials where data are collected prospectively until an event of interest occurs. In such studies, the time-to-event (TTE) is usually the outcome variable. This outcome variable is called failure time, event time or survival time. To determine the relationship between covariates and survival time, survival analysis techniques are used [14]. Survival data analysis refers primarily to a set of methods for analysing survival data to study the distribution of lifetimes or survival times (i.e. the time until the occurrence of a specified event of interest) [14]. The event of interest can be death, development of a disease, relapse of disease after treatment, etc. and the survival time may be measured in days, weeks, bi-weekly, months, quarterly, years, etc. For example, in this thesis, the event of interest is colorectal cancer (CRC) and the survival time is the time (in years) until an individual develops CRC. So why not use ordinary linear regression (OLR) modelling techniques to model survival time as a function of covariates? Three reasons [14-15]: (i) survival times are always positive values and OLR may not be the ideal choice of a method unless the survival times are transformed in such a way that this restriction is eliminated, (ii) the distribution of survival data is often asymmetric (skewed) and thus techniques including OLR which are based on the Gaussian distribution cannot be used directly, and (iii) most importantly, OLR cannot efficiently handle survival data due the presence of censored observations [14].

In the aforementioned techniques, survival times are independent and the variance of the regression coefficients obtained from the Cox proportional hazards (PH) regression model are consistent [15]. In section 4.6.2, survival analysis techniques for analysing cross-sectional independent data are provided. Terminologies commonly used in survival analysis are defined in sections 4.6.2.1 through to 4.6.2.4. Details of the Cox proportional hazards (PH) regression model are provided in section 4.6.2.5. The discrete PH model and discrete survival with logit and complementary log-log links are provided in sections 4.6.2.5.1 and

4.6.2.5.2 respectively. The statistical application of survival analysis techniques to answer objective 2 of this thesis is provided in section 4.6.3.

4.6.2 Survival Analysis Techniques for Cross-sectional Independent Data

Survival analysis techniques are used to analyse time-to-event data. Collecting survival data for incident cases often span over (relatively) longer study periods in order to gather enough cases for meaningful analysis [154]. In studies that collect survival data, groups of subjects are followed over a specified period of time to determine the occurrence of an event of interest. For example, a disease-free cohort is followed over a period of time to determine the occurrence of colorectal cancer. In survival analysis, several terminologies including survival time, probability density function (pdf) of the survival time, censoring, survival function, hazard function, hazard rate, and hazard ratio are often used. A description of these terminologies is provided in the following sections and the details of which are given by kleinbaum and Klein [14] and Cox [15].

4.6.2.1 Survival time

In survival analysis, the time until the occurrence of a specified event is called survival time (T) [14-15]. As stated earlier (see section 4.6.1), survival time may be measured in days, weeks, bi-weekly, months, quarterly, or years until the event occurs which is typically called failure [14]. There are unique features of survival time [14-15]. First, survival times are always non-negative and usually have skewed distributions [14]. For instance, in a study evaluating the time to relapse of colorectal cancer among high-risk patients, the majority of failures (relapses) are likely to occur earlier in the follow-up with relatively few failures occurring later. However, in a study evaluating the time to death, the majority of failures (deaths) may occur later in the follow-up. Statistical methods based on the Gaussian distribution cannot be used to analyse the survival time [2]. Non-parametric methods could however be used [12, 14]. Complete information (actual time to event data) is often not available for each subject and those who enrol earlier in the study may be followed for a longer period while those who enrol later may be followed for a shorter period (and the opposite may be true too) probably because study participants may have moved, withdrawn from the study or died (assuming the outcome of interest is not death) [14]. Each of these scenarios will result in incomplete data for some participants and the true survival/failure time will not be known because the study ended or because the event of interest does not occur when the subject dropped out of the study. This results in survival times for some for subjects that are greater than their last recorded/observed follow-up time. These times are termed censored times [14].

4.6.2.2 Censoring

A unique feature that distinguishes survival data from data collected in other fields of statistics is the presence of censoring in the former. Censoring occurs when there is incomplete knowledge about the survival time of some subjects [14-15]. This may arise because some subjects are lost to follow-up or withdraw from the study without experiencing the event of interest or the event does not occur for some subjects at the end of the study period. There exist different types of censoring mechanisms in survival data analysis and these include right censoring, left censoring, and interval censoring [14-15, 50]. Right censoring occurs when an individual does not experience the event of interest before the end of the study or dropped out before the event occurs and so their last observed/recorded follow-up time is less than their time to the event [14, 50]. Right censoring is sometimes called administrative censoring [14]. For left censoring, the individual would have already experienced the event of interest at the time they joined the study [14, 50]. Unlike right and left censoring, interval censoring occurs when the specific time the event occurred is not known, but the first and last time the experimental unit was measured with and without the event respectively are available [50]. For the survival analysis conducted in this thesis, we assumed that interval censoring is non-informative i.e. the time of censoring (of survival time) is independent of the time of occurrence of CRC that otherwise would have been observed given the covariates used in this thesis.

Mathematically, let T_i be a non-negative random variable representing survival time and C_i be the random variable representing censorship for the i^{th} person in a sample of n subjects. If $X_i = min(T_i\,, C_i\,)$, then we define an indicator variable δ_i such that;

$$\delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i \\ 0 & \text{if } T_i > C_i \end{cases}$$

where $\delta_i = 1$ indicates that the i^{th} individual is uncensored and the event of interest has occurred and $\delta_i = 0$ indicated censorship for the i^{th} individual and the event of interest has not occurred [50].

4.6.2.3 Functions of Survival time (adopted and modified from [2] and [14])

Survival function [$S(t)$]: This is the probability that an individual survives beyond some specified time t. Mathematically, the survival time is defined as;

$$S(t) = \mathcal{P}r(T \geq t), 0 \leq t \leq \infty.$$

Some theoretical properties of the survival function $S(t)$ include: As $t \in (0, \infty)$,

1. The survival function is monotonically non-increasing.
2. At time $= 0,\ S(t) = 1$. This means that the probability of surviving beyond time 0 is 1. In other words, the no event can occur before the study starts and everybody survives (with probability 1) at time $t = 0$.
3. When $t = \infty, S(t) = S(\infty) = 0$. Theoretically, this means that as the study time increases to infinity, the survival curve reaches zero (0). In other words, if the study time was to be increased infinitely, all subjects will eventually experience the outcome (i.e. failure will occur for all subjects).

The survival function $S(t)$ is also referred to as the cumulative survival rate, the graph [$S(t)$ on the y-axis and time t on the x-axis)] of which is used to determine the 50th percentile, also known as the median survival time.

***Probability density function (pdf)* [$\boldsymbol{f(t)}$]:** The pdf of the survival time T is the distribution of survival times which describes the probability of failing within an infinitesimally small interval $[t, \Delta t]$ per unit time. Mathematically, $f(t)$ is defined as the limit of the probability that an individual experiences the event of interest in the interval $[t, \Delta t]$ per unit width Δt and we write

$$f(t) = \lim_{\Delta t \to 0} \frac{\mathcal{P}r(t \leq T \leq t + \Delta t)}{\Delta t}$$

Some theoretical properties of the $f(t)$ include:

1. The density function $f(t)$ is non-negative. That is $f(t) \geq 0, \forall\, t \geq 0$ and $f(t) = 0$ for $t < 0$.
2. The total area under the density curve is 1.

Hazard function $[\lambda(t)]$**:** This is the conditional failure rate. It is also called hazard rate, failure rate or force of mortality. The hazard function $[\lambda(t)]$ describes the probability of experiencing the outcome in a small interval of time $[t, \Delta t]$ assuming an individual survived to the beginning of the specified interval. Mathematically, $\lambda(t)$ is the limit of the probability of failing within a small interval per unit time given that an individual survived to the beginning of the interval and we write;

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathcal{P}r(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function $\lambda(t)$ may reduce, increase, or remain unchanged and has no upper limit. It is not a probability value (so it can be more than 1) and also depends on the unit of time. Appendix C summarises the computational definitions for these functions of survival time when T either continuous or discrete.

4.6.2.4 Hazard Ratio (HR)

The hazard ratio describes the instantaneous relative hazard/risk of an event occurring per unit time for a subject when a specified risk factor is present compared to hazard/risk of the event occurring per unit time for another subject when the risk factor is not present when both subjects have survived to a specified time t. For a continuous explanatory variable, the hazard rate is defined as:

$$\frac{\lambda(t|x_i + \Delta)}{\lambda(t|x_i)} = e^{\beta_i \Delta},$$

whereas for a categorical explanatory variable, the hazard rate is defined as:

$$\frac{\lambda(t|x_i = 1)}{\lambda(t|x_i = 0)} = e^{\beta_i}.$$

A numeric value of the hazard ratio indicates the relative hazard/risk reduction achieved when the specified explanatory variable is present as compared to the hazard/risk reduction when the explanatory variable is not present.

4.6.2.5 The Cox Proportional Hazards (PH) Regression Model

Cox [15] introduced a large family of models that are directly focused on the hazard function. The simplest member of this family of models is the proportional hazards model. In the proportional hazards model, the hazard at time t for the i^{th} subject with a set of covariates $\boldsymbol{X}_i$ (excluding the constant term) is assumed to be $\lambda_i(t|\boldsymbol{X}_i)$ and defined as follows:

$$\lambda_i(t|\boldsymbol{X}_i) = \lambda_0(t)\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta}) = \exp(\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p) \ldots\ldots\ldots\ldots\ldots\ldots..(4.21)$$

where $\lambda_0(t)$ represents the baseline hazard function and describes the risk associated with individuals with $\boldsymbol{X}_i = 0$ (i.e. the reference group), and $\exp\{\boldsymbol{X'}_i\boldsymbol{\beta}\}$ is the relative risk measuring the proportionate change (i.e. increase or reduction) in the risk associated with $\boldsymbol{X}_i$. This change in risk is the same at all times t.

To juxtapose the two risk components, consider a two-sample scenario where X_i is a binary dummy variable serving only to distinguish between the two groups [say groups zero (0) and one (1)]. Then the appropriate model is;

$$\lambda_i(t|\boldsymbol{X}_i) = \begin{cases}\lambda_0(t) & \text{if}\,\boldsymbol{X_i} = 0\\ \lambda_0(t)\gamma & \text{if}\,\boldsymbol{X_i} = 1\end{cases} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4.22)$$

where $\lambda_0(t)$ represents the risk associated with group 0 at time t and $\gamma = \exp(\beta)$ represents the ratio of the risk associated with group one relative to the risk associated with group zero at any time t. In particular, when $\gamma = 1$ (or $\beta = 0$), the risks are the same for the two groups. However, assuming age is the covariate in question, when $\gamma = 2$ (or $\beta = 0.6931$), the risk associated with a person in group one at any given age) is two times the risk associated with a person in group zero who has the same age. The model in equation (4.22) separates the effects of the covariates $\boldsymbol{X}_i$ from the effects of time. Simple manipulation of equation (4.21) reveals that the proportional hazards model is an additive model for the log of the hazards. To show this, we take the natural logarithms of the proportional hazards models in equation (4.21) as;

$$\begin{aligned}\ln[\lambda_i(t|\boldsymbol{X}_i)] &= \ln\,[\lambda_0(t)\exp\{\boldsymbol{X'}_i\boldsymbol{\beta}\}]\\ &= \ln[\lambda_0(t)] + \boldsymbol{X'}_i\boldsymbol{\beta}\\ &= \alpha_0(t) + \boldsymbol{X'}_i\boldsymbol{\beta} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(4.23)\end{aligned}$$

where $\alpha_0(t) = \ln[\lambda_0(t)]$ is the log of the baseline hazard. Note that $\alpha_0(t)$ is independent of the X's and thus show that the effects of the covariates are constant/same at all times t. To compare the hazards rates of two individuals with two covariate vectors $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ respectively, we use the hazards ratio defined as:

$$\frac{\lambda_1(t|\boldsymbol{X}_i)}{\lambda_2(t|\boldsymbol{X}_j)} = \frac{\lambda_0(t)\exp\,(X_i\beta)}{\lambda_0(t)\exp\,(X_j\beta)} = \exp\left[\beta\left(X_i - X_j\right)\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(4.24)$$

where $\lambda_1(t|\boldsymbol{X}_i)$ and $\lambda_2(t|\boldsymbol{X}_j)$ are the hazards rates for the two individuals. Equation (4.24) assumes that hazard rates for any two individuals are proportional and does not depend on time;

Closely related to the proportional hazards in equation (4.21) is the cumulative hazards which are also proportional (see equation 4.25) [14-15]. The cumulative hazards, $\Lambda_i(t|X_i)$, is derived from proportional hazards by integrating the latter over $[0, t]$ as follows:

$$\begin{aligned}\Lambda_i(t|\mathrm{X}_i) &= \int_0^t \lambda_0(u)\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})du\\ &= \exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})\int_0^t \lambda_0(u)du\\ &= \exp(\boldsymbol{X'}_i\boldsymbol{\beta})\,\Lambda_0(t), \ldots\ldots\ldots\ldots\ldots\ldots.\ldots\ldots\ldots\ldots\ldots\ldots.(4.25)\end{aligned}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(u)du$ is the cumulative baseline hazard.

One of the advantages of the Cox model lies in its ability to incorporate coefficients and covariates that change over time [15]. Time-dependent coefficients and time-dependent covariates are two extensions of the Cox model used to model non-proportional hazards [14]. Equations (4.26) and (4.27) describe time-dependent coefficient and time-dependent covariate models respectively. Equation (4.28) combines both equations (4.26) and (4.27) to produce a generic version of the hazard rate model which models both the time-dependent coefficient and time-dependent covariate simultaneously [14].

$$\lambda_i(t|\boldsymbol{X}_i) = \lambda_0(t)\exp[\boldsymbol{X}'_i\boldsymbol{\beta}(t)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{4.26}$$

$$\lambda_i(t|\boldsymbol{X}_i(t)) = \lambda_0(t)\exp[\boldsymbol{X}'_i(t)\boldsymbol{\beta}] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{4.27}$$

$$\lambda_i(t|\boldsymbol{X}_i(t)) = \lambda_0(t)\exp[X'_i(t)\boldsymbol{\beta}(t)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \tag{4.28}$$

where $X'_i(t)$ is a vector of time-dependent covariates for the i^{th} subject at time t and $\boldsymbol{\beta}(t)$ is a time-dependent vector of coefficients.

4.6.2.5.1 The Discrete Proportional Hazards (PH) Model

In survival analysis, the outcome of interest is the time until the occurrence of an event of interest. In the Saskatchewan Rural Health Study (SRHS) dataset, the exact time to the occurrence of colorectal cancer (CRC) is not known. However, only the time interval in which the disease could have occurred is known. This becomes a discrete process which results in interval censored variables. Survival data collected through a discrete process may be called interval censored data and can be analyzed using discrete-time modeling techniques for a determination of the hazard or failing or surviving.

Let T_i, a non-negative random variable taking values $t_1, t_2, \ldots$ be the time until the occurrence of CRC. Then the discrete-time hazard function p_{t_i}, is the probability that the i^{th} subject is diagnosed with CRC in a specified interval given that the individual was not diagnosed with CRC before the start of the period in question [188]. That is the conditional probability of failing in the interval given that the individual survived till the start of the period. The discrete-time hazard function [188] is given as;

$$p_{t_i} = \mathcal{P}r\left[T_i \in \left(t_i^0, t_i^f\right) \middle| T_i \geq t_i^0\right] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (4.29)$$

where t_i^0 is the time of origin (in this thesis, it is the start of the baseline survey in 2010) and t_i^f is the endpoint time (in this thesis, it is the end of the follow-up survey in 2014). p_{t_i} is a discrete-time approximation to the continuous-time function (described in section 4.6.2.3). It is not too difficult to observe from equation (4.29) that in order for an individual can survive to time t_i^f, the person must first survive through t_1, then the person must survive to t_2 given that he/she has survived through t_1, and so on and so forth, finally surviving through t_i^{f-1} given that he/she has survived up to that point.

We obtain the baseline hazard for the time interval $\left(t_i^0, t_i^f\right)$ by manipulating the discrete-time hazard function p_{t_i} using the complement of the hazard function $(1 - p_{t_i})$. To show this, let λ_{ij} be a discrete hazard which represents the probability of the i^{th} subject failing in interval $(j, j+1)$ given that the person survived to the start of j and we write;

$$\begin{aligned}
\lambda_{ij} &= 1 - p_{t_i} \\
&= 1 - \mathcal{P}r\left[T_i \in \left(t_i^0, t_i^f\right) \middle| T_i \geq t_i^0\right] \\
&= 1 - \exp\,[-\int_{t_i^0}^{t_i^f} \lambda_i(t|\boldsymbol{X}_i)\,dt] \\
&= 1 - \exp\left\{-\int_{t_i^0}^{t_i^f} [\lambda_0(t)\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})]\,dt\right\} \\
&= 1 - \exp\left\{-\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}^{\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})}
\end{aligned}$$

$$1 - \lambda_{ij} = \exp\left\{-\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}^{\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})}.$$

Taking logs of both sides of the last equation, we have

$$\log\,(1-\lambda_{ij}) = \log\left\{\exp\,\left\{-\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}^{\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})}\right\}$$

$$-\log\,(1-\lambda_{ij}) = \exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt$$

Taking logs of both sides again, we have

$$\log\bigl[-\log\bigl(1-\lambda_{ij}\bigr)\bigr] = \log\left\{\exp\,(\boldsymbol{X'}_i\boldsymbol{\beta})\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}$$

$$= \log\,(\exp(\boldsymbol{X'}_i\boldsymbol{\beta}) + \log\left\{\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}$$

$$= (\boldsymbol{X'}_i\boldsymbol{\beta}) + \log\left\{\int_{t_i^0}^{t_i^f} \lambda_0(t)\,dt\right\}\ldots\ldots\ldots\ldots\ldots\ldots.(4.30)$$

where the quantity $\int_{t_i^0}^{t_i^f}\lambda_0(t)\,dt$ is the baseline risk for the interval $\left(t_i^0, t_i^f\right)$. If $\lambda_0(t)$ does not vary rapidly over $\left(t_i^0, t_i^f\right)$, then $\int_{t_i^0}^{t_i^f}\lambda_0(t)\,dt \cong \overline{\lambda_0}(t_i^f - t_i^0)$, where $\overline{\lambda_0}$ represents the mean baseline hazard. Given this, equation (4.30) is be modified as;

$$\log\bigl[-\log\bigl(1-\lambda_{ij}\bigr)\bigr] = (\boldsymbol{X'}_i\boldsymbol{\beta}) + \log\bigl[\overline{\lambda_0}\bigl(t_i^f - t_i^0\bigr)\bigr]$$

$$= (\boldsymbol{X'}_i\boldsymbol{\beta}) + \log\bigl(\overline{\lambda_0}\bigr) + \log\bigl(t_i^f - t_i^0\bigr)$$

$$= (\boldsymbol{X'}_i\boldsymbol{\beta}) + \alpha_0 + \log\bigl(t_i^f - t_i^0\bigr), \ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4.31)$$

where $\alpha_0 = \log(\overline{\lambda_0})$ (i.e. the log of the mean baseline hazard). Equation (4.31) is used in studies with independent non-recurrent events. However, it (equation (4.31)) is modified to account for the repeated observations in longitudinal studies where separate measurements exist for the i^{th} subject at the different time points [188]. The longitudinal version of equation (4.31) is given as;

$$\log[-\log(1-\lambda_{ij})] = (\boldsymbol{X'}_{ij}\boldsymbol{\beta}) + \alpha_{0j} + \log(t_{ij}^{f} - t_{ij}^{0}) \dots\dots\dots\dots\dots\dots\dots\dots\dots. (4.32)$$

where $t_{ij}^{f} - t_{ij}^{0}$ represents the risk time and α_{0j} quantifies any variation among the baseline hazards across the different time points (in our case, across the baseline and follow-up surveys) [188].

4.6.2.5.2 Discrete-Time Survival Analysis (DTSA) and the Logit Link

Cox [15] extended proportional hazards modeling techniques for discrete time by working with the conditional odds of failing at each time t_j given that an individual survived to this time point. He proposed the model

$$\frac{\lambda(t_j|\boldsymbol{X}_i)}{1-\lambda(t_j|\boldsymbol{X}_i)} = \frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}\exp(\boldsymbol{X'}_i\boldsymbol{\beta}), \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots. (4.33)$$

where $\lambda(t_j|X_i)$ represents the hazard at time t_j for the i^{th} subject with a set of covariate values $\boldsymbol{X}_i$, $\lambda_0(t_j)$ represents the baseline hazard at time t_j, and $\exp(\boldsymbol{X'}_i\boldsymbol{\beta})$ measures the relative risk associated with X_i. We obtain a logit model for the hazard of failing at each time t_j given that an individual survived to this time point by applying logs to (4.33) as;

$$\log\left\{\frac{\lambda(t_j|\boldsymbol{X}_i)}{1-\lambda(t_j|\boldsymbol{X}_i)}\right\} = \log\left\{\left\{\frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}\right\}\exp(\boldsymbol{X'}_i\boldsymbol{\beta})\right\}$$

$$= \log\left\{\frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}\right\} + (\boldsymbol{X'}_i\boldsymbol{\beta})$$

$$\text{logit}[\lambda(t_j|X_i)] = \alpha_j + (\boldsymbol{X'}_i\boldsymbol{\beta}) \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots.. (4.34)$$

where $\alpha_j = \log\left\{\frac{\lambda_0(t_j)}{1-\lambda_0(t_j)}\right\}$ represents the logit of the baseline hazard and $\boldsymbol{X'}_i\boldsymbol{\beta}$ is the effect of the covariates $\boldsymbol{X}_i$ on the logit of the hazard. The model specified in equation (4.34) considers time as discrete by introducing the parameter α_j for each possible failure time t_j. The regression coefficients $\boldsymbol{\beta}$ are interpreted just as in a classical logistic regression model.

4.6.2.5.3 Discrete-Time Survival Analysis (DTSA) and the Complementary Log-Log (C-Log-Log) Link

An alternate extension of the Cox's proportional hazard model to include discrete time lie in the expression of the survival function. In a proportional hazard framework, the survival function is expressed as;

$$S(t_j|\boldsymbol{X}_i) = S_0(t_j)\exp(\boldsymbol{X'}_i\boldsymbol{\beta}),$$

where $S(t_j|\boldsymbol{X}_i)$ represents the probability the i^{th} subject with a set of covariate values $\boldsymbol{X}_i$ will survive to time t_j and $S_0(t_j)$ represents the baseline survival function. By expressing the survival function in terms of the hazard and applying a similar approach to that used in deriving equation (4.30), we obtain an analogous relationship for the for the complement of the hazard function as;

$$1 - \lambda(t_j|\boldsymbol{X}_i) = \{1 - \lambda_0(t_j)\}^{\exp(\boldsymbol{X'}_i\boldsymbol{\beta})},$$

and further, obtain the hazard for the i^{th} subject failing at time point t_j as

$$\lambda(t_j|\boldsymbol{X}_i) = 1 - \{1 - \lambda_0(t_j)\}^{\exp(\boldsymbol{X'}_i\boldsymbol{\beta})}.$$

The complementary log-log (c-log-log) transformation is used to linearize $\lambda(t_j|\boldsymbol{X}_i)$ as a function of parameters to obtain the model

$$\log\left\{-\log\{1 - \lambda(t_j|\boldsymbol{X}_i)\}\right\} = \alpha_j + \boldsymbol{X'}_i\boldsymbol{\beta}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(4.35)$$

where $\alpha_j = \log\left\{1 - \{1 - \lambda_0(t_j)\}\right\}$ represents the c-log-log transformation of $\lambda_0(t_j)$, the baseline hazard. The Cox proportional model can, therefore, be fitted to discrete survival data by using a binary response GLM with binomial error structure and a c-log-log link. This means that, under non-informative sampling, both the binomial likelihood as well as the discrete time survival likelihood hold for the logit and complementary log-log links.

The choice of a link function in survival analysis depends on how time is measured. If time is measured as a discrete variable, then the logit link is more appropriate. However, if time is measured as a continuous variable that it truly is, then c-log-log link is a better choice.

4.6.3 Objective Two (2): Survival Analysis

Cox's proportional hazard (PH) model was used to determine the adjusted incidence of CRC in rural Saskatchewan. The crude incidence rate was computed using the cumulative incidence formula (see section 4.6.3.1.1). To study the effects of individual and contextual factors as well as covariates on CRC incidence, the proportional hazards model (see section 4.6.2.5) is used.

4.6.3.1 Statistical Application: Objective Two (2)

4.6.3.1.1 Crude Incidence of CRC

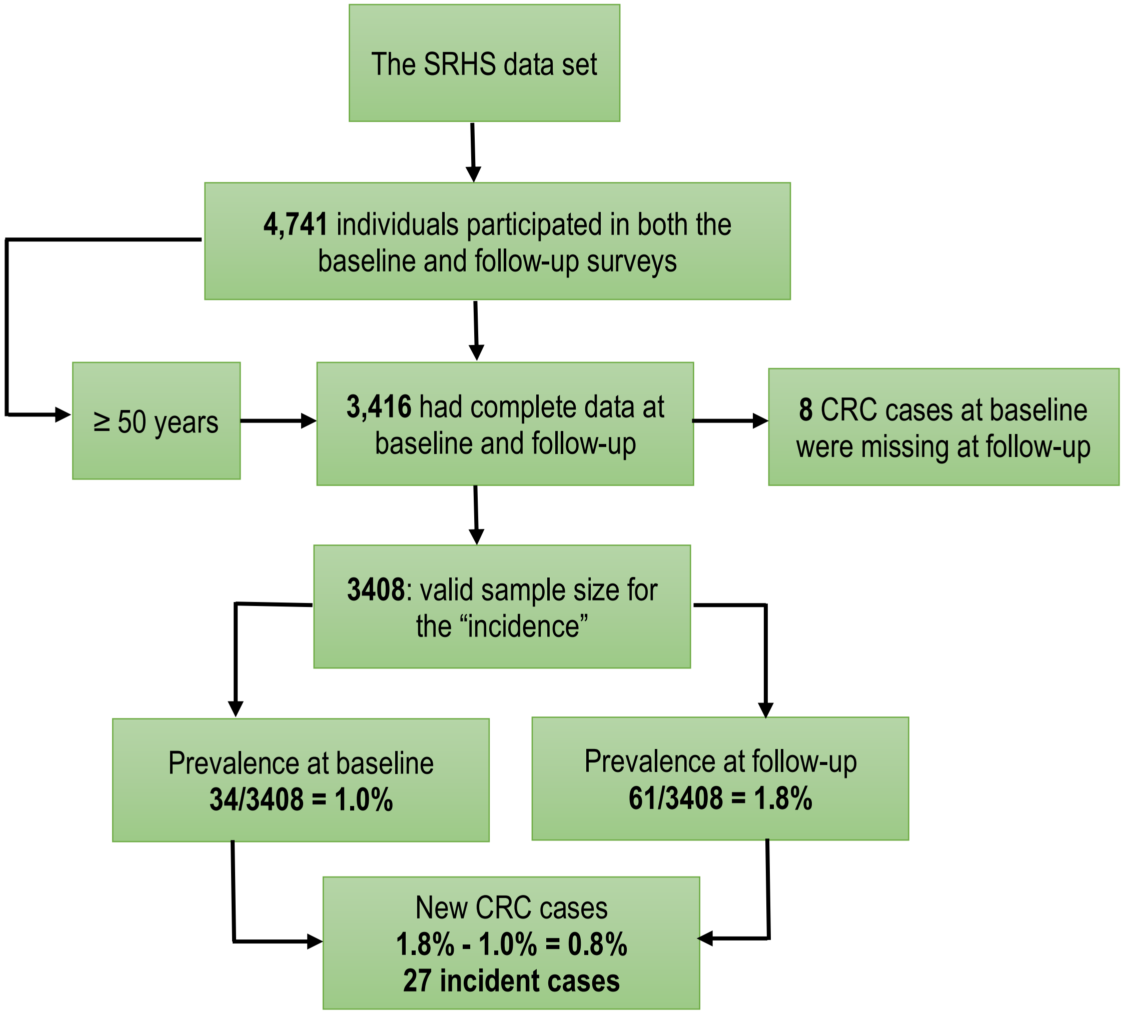
In this thesis, we adopt the definition of incidence as "the number of new events of a specific disease during a specified period of time in a specified population" [155]. The incidence rate is "the rate at which new events, or new cases, occur in a specified time in a defined population that is at risk of experiencing the condition or event" [155]. The incidence rate was computed using the formula below;

$$Incidence\ rate = \frac{\#\ of\ new\ cases\ in\ specified\ period}{\#\ of\ people\ at\ risk\ in\ the\ specified\ period} \text{ x } 100 \ldots\ldots\ldots\ldots\ldots (**)$$

Using this formula, we calculated the crude incidence rate, also known as the cumulative incidence rate [155] of CRC as follows;

Figure 4.2 is a schematic representation of the selection of new CRC cases and a valid samples size for incidence analysis.

Figure 4. 2 Illustrative representation of selecting new CRC cases in the SRHS data set


The SRHS data set
4,741 individuals participated in both the baseline and follow-up surveys
≥ 50 years
3,416 had complete data at baseline and follow-up
8 CRC cases at baseline were missing at follow-up
3408: valid sample size for the "incidence"
Prevalence at baseline 34/3408 = 1.0%
Prevalence at follow-up 61/3408 = 1.8%
New CRC cases 1.8% - 1.0% = 0.8% 27 incident cases

4.6.3.1.2 Adjusted Incidence of CRC

Before commencing the data analysis to determine risk factors for CRC incidence, the data set needed to be rearranged to be able to perform survival analysis. Two (2) new variables were created and denoted as *incidence* and *time-to-event (CRC)* variables. The *incidence* variable was a dichotomous variable assuming values 0 or 1. When an individual is diagnosed with CRC during follow-up, the *incidence* variable was assigned a value of 1 and 0 otherwise. On the other hand, the *time-to-event* variable represents the time an individual is followed until he/she develops CRC. The baseline survey will be the time of origin and the follow-up will be the time-to-an-event (i.e. CRC incident cases). In this thesis, the exact time of the occurrence of colorectal cancer (CRC) is not known. However, the time interval in which the disease could have occurred is known. This becomes a discrete process which results in interval censored variables with follow-up time for censored cases being four (4) years. Censoring will occur when an individual was not diagnosed with CRC at the end of follow-up.

The PH model was used to determine risk factors for the incidence of CRC. The primary aim of using the PH model was to identify the most parsimonious model to describe CRC incidence between the baseline and follow-up surveys. Standard model building techniques were used in the selection of the final multivariable model [26]. Crude hazard rates were calculated in the univariable analysis and variables with p-value <0.25 were candidates for the multivariable model. Biologically relevant such as gender and BMI were retained in the multivariable analysis even though they were not significant in the univariable analysis. The proportional hazards model assumptions as well as the model fit were tested using Schoenfeld residuals [14].

Before fitting the proportional hazards model, the STSET command in STATA was used to declare the CRC incidence data as survival-time data [14]. To fit the proportional hazards model, the STCOX command was used [14]. The STCOX command fits a PH model based on the maximum likelihood theory [189]. Clustering effects of more than one individual in a household was accounted using the clustered estimator *VCE(CLUSTER HOUSEID)* command, where *HOUSEID* is the within subject variable in this study. The STPHTEST command was used to test the proportional hazards model assumption.

CHAPTER 5 – RESULTS

5.1 Introduction

The primary goal of this thesis was to determine the prevalence and incidence of self-reported doctor-diagnosed CRC and associated risk factors in rural Saskatchewan using appropriate statistical methodology models to analyze survey data with a binary outcome. The remainder of this chapter is organized follows; Section 5.2 describes the sample population in this study while Section 5.3 provides some descriptive statistics. Research questions one and two are answered in Sections 5.4 and 5.5 respectively. Sections 5.6 summarizes the methods used to analyze the data used in this thesis.

5.2 Sample population

The SRHS contains information of 8,261 participants nested within 4,624 households at baseline and 4,867 participants nested within 2,797 households at follow-up. A subset of the SRHS data was extracted and used for the analysis contained in this thesis. A total of 5,599 individuals at baseline and 3,933 at follow-up aged 50 or older were included in the present analysis because (1) this is the population we considered at risk and (2) the youngest person diagnosed with CRC in this study was 50 years old.

5.3 Descriptive Characteristics

Table 5.1 gives the population characteristics stratified by farm and non-farm residence status and survey. There was a reduction in the number of participants from baseline to follow-up. This reduction occurred in both the farm and non-farm sub-categories. The majority (44.6%) of farming residents were in their 50s as compared to 33.6% of non-farm residents at baseline. On the other hand, a majority (36.1%) of non-farm residents were aged 70 years or older as compared to 22.6% of farm residents. A similar trend was observed at the end of the follow-up survey (Table 5.1). Approximately 30% of both farm and non-farm residents were of normal weights and obese respectively both at baseline and follow-up. About the same percentage of both farm and non-farm residents were over weighted (≈ 43%). This trend continued at follow-up (farm residents: 41.5% vs. non-farm-residents: 42.3%) (Table 5.1). At baseline, more non-farm residents (34.8%) completed Grade 12 or higher as compared to farm residents (32.1%). The study population contained an approximately equal number of males and females both at baseline as well as at follow-up (Table 5.1).

Shallow well was the source of household water supply to a majority (34.6%) of farm residents as compared to non-farm residents (9.7%) at baseline. This trend persisted during the four-year follow-up period (farm residents: 31.7% vs. non-farm residents: 9.8%). More farm households (27.8%) use bottled water as compared to non-farm households (26.2%). A similar trend was observed among farm households (29.9%) as compared to non-farm households (26%) at follow-up. The percentage of non-farm residents (88.2%) were more than twice higher than the percentage of farm residents (39.2%) who reported natural gas as their major source of fuel to their households at baseline. A similar two-fold percentage occurred at follow-up (Table 5.1). More non-farm residents (15.6%) smoked in their homes when compared to farm residents (12.4): an observation which sufficed at follow-up (non-farm residents – 11.1% vs. farm residents – 8.8%. About 9% of farm residents were current smokers as compared to 11.4% of non-farm residents.

We also observed that a higher number of non-farm residents consume alcohol daily than farm residents (Table 5.1). About 80% of non-farm residents ever lived on a farm while about 70% of them lived on a farm in their first year of life (Table 5.1). Only 10.2% of farm residents had a previous history of cancer as compared to 11.7% of non-farm residents. About 30% of both farm and non-farm residents had a family history of cancer (i.e. either the father, mother or sibling ever had a previous history of cancer) (Table 5.1). More farm residents were exposed to diesel and welding fumes, fungicides, grain dust, pesticides, livestock, molds, oil/gas well fumes, stubble smoke, and wood dust at their workplaces than non-farm residents (Table 5.1).

Before data analysis, race/ethnic group was studied. However, preliminary data analysis revealed that 98% of the study population in this thesis were Caucasian while only 2% were non-Caucasian. A decision was made to exclude race from all analysis.

Table 5. 1 Population Characteristics stratified of rural residents (≥ 50 years) by farm and non-farm residence status, Saskatchewan

Description	**Baseline** (N = 5599)			**Follow-up** (N = 3933)		
	Farm	**Non-farm**	**Missing**	**Farm**	**Non-farm**	**Missing**
	N = 2339	**N = 3212**	**N = 48**	**N = 1676**	**N = 2236**	**N = 21**
	n (%)	n (%)	n (%)	n (%)	n (%)	n (%)
CONTEXTUAL FACTORS						
Socioeconomic			670 (12.0)			464 (11.8)
Some money	1267 (61.6)	1725 (60.1)		1046 (71.6)	1335 (66.5)	
Just enough money	368 (17.9)	640 (22.3)		254 (17.4)	411 (20.5)	
Not enough money	422 (20.5)	507 (17.7)		161 (11.0)	262 (13.0)	
Quadrant			52 (0.9)			98 (2.5)
South West	372 (15.9)	656 (20.4)		257 (15.6)	448 (20.5)	
South East	439 (18.8)	720 (22.4)		317 (19.3)	476 (21.70	
North East	856 (36.6)	867 (27.0)		587 (35.7)	603 (27.5)	
North West	672 (28.7)	965 (30.1)		485 (29.5)	662 (30.2)	
Environmental:						
Household smoking			79 (1.4)			21 (0.5)
Yes	288 (12.4)	499 (15.6)		147 (8.8)	248 (11.1)	
No	2043 (87.6)	2690 (84.4)		1529 (91.2)	1988 (88.9)	
Household structure:						
Water source			48 (0.9)			87 (2.2)
Bottled Water	651 (27.8)	840 (26.2)		497 (29.9)	569 (26.0)	
Deep well water (more than 100ft)	524 (22.4)	854 (26.6)		348 (21.0)	566 (25.9)	
Shallow well water (less than 100ft)	809 (34.6)	311 (9.7)		527 (31.7)	214 (9.8)	
Other source	355 (15.2)	1207 (37.6)		289 (17.4)	836 (38.3)	
Fuel source – Natural gas			62 (1.1)			31 (0.8)
Yes	916 (39.2)	2824 (88.2)		656 (39.2)	1975 (88.6)	
No	1419 (60.8)	378 (11.8)		1017 (60.8)	254 (11.4)	
Mildew odor or musty smell in home			227 (4.1)			124 (3.2)
Yes	429 (19.1)	352 (11.3)		305 (18.7)	311 (14.3)	
No	1819 (80.9)	2772 (88.7)		1323 (81.3)	1870 (85.7)	

INDIVIDUAL FACTORS						
Smoking Status			85 (1.5)			56 (1.4)
Current Smoker	205 (8.8)	364 (11.4)		125 (7.6)	219 (9.9)	
Ex-smoker	840 (36.1)	1419 (44.5)		625 (37.8)	986 (44.4)	
Never smoker	1283 (55.1)	1403 (44.0)		905 (54.7)	1017 (45.8)	
Alcohol consumption			75 (1.3)			48 (1.2)
Never	442 (19.0)	718 (22.5)		319 (19.1)	416 (18.8)	
Less than once a month	492 (21.1)	704 (22.0)		338 (20.3)	462 (20.8)	
At most 2-3 times a month	629 (27.0)	715 (22.4)		411 (24.6)	514 (23.2)	
At most 2-3 times a week	516 (22.2)	645 (20.2)		389 (23.3)	491 (22.1)	
Everyday	250 (10.7)	413 (12.9)		211 (12.6)	334 (15.1)	
Physical activity			244 (4.4)			117 (3.0)
Yes	1280 (56.5)	1781 (57.7)		856 (52.2)	1213 (55.7)	
No	986 (43.5)	1308 (42.3)		783 (47.8)	964 (44.3)	
Diabetes			135 (2.4)			131 (3.3)
Yes	213 (9.2)	448 (14.2)		118 (389)	271 (12.5)	
No	2097 (90.8)	2706 (85.8)		1517 (92.8)	1896 (87.5)	
Early life-exposures:						
Ever-lived on a farm			65 (1.2)			25 (0.6)
Yes	2280 (97.5)	2554 (79.9)		1660 (99.1)	1801 (80.7)	
No	58 (2.5)	642 (20.1)		15 (0.9)	432 (19.3)	
Lived on a farm in first year of life			116 (2.1)			52 (1.3)
Yes	1917 (82.8)	2146 (67.7)		1391 (83.3)	1537 (69.5)	
No	397 (17.2)	1023 (32.3)		278 (16.7)	675 (30.5)	
Hereditary:						
Familial History of cancer:						
Personal history of cancer			129 (2.3)			132 (3.4)
Yes	237 (10.2)	370 (11.7)		132 (8.1)	223 (10.3)	
No	2077 (89.8)	2786 (88.3)		1501 (91.9)	1945 (89.7)	
Father ever had cancer			596 (10.6)			396 (10.1)
Yes	724 (33.7)	915 (32.0)		507 (33.1)	659 (32.9)	
No	1423 (66.3)	1941 (68.0)		1026 (66.9)	1345 (67.1)	
Mother ever had cancer			504 (9.0)			325 (8.3)

Yes	580 (26.6)	856 (29.4)		419 (26.8)	614 (30.0)	
No	1600 (73.4)	2059 (70.6)		1142 (73.2)	1433 (70.0)	
Sibling(s) ever had cancer			978 (17.5)			648 (16.5)
Yes	550 (28.0)	889 (33.5)		358 (25.2)	553 (29.7)	
No	1417 (72.0)	1765 (66.5)		1064 (74.8)	1310 (70.3)	
Occupational Exposures:						
At work, ever exposed to:						
Asbestos dust			191 (3.4)			150 (3.8)
Yes	132 (5.7)	240 (7.8)		92 (5.6)	160 (7.5)	
No	2192 (94.3)	2844 (92.2)		1547 (94.4)	1984 (92.5)	
Diesel fumes			191 (3.4)			150 (3.8)
Yes	1706 (73.4)	1537 (49.8)		1220 (74.4)	1075 (50.1)	
No	618 (26.6)	1547 (50.2)		419 (25.6)	1069 (49.9)	
Fungicides (to treat grain)			191 (3.4)			150 (3.8)
Yes	1088 (46.8)	851 (27.6)		785 (47.9)	598 (27.9)	
No	1236 (53.2)	2233 (72.4)		854 (52.1)	1546 (72.1)	
Grain dust			191 (3.4)			150 (3.8)
Yes	2045 (88.0)	1773 (57.5)		1439 (87.8)	1225 (57.1)	
No	279 (12.0)	1311 (42.5)		200 (12.2)	919 (42.9)	
Pesticides (to kill plants and insects)			48 (0.9)			21 (0.5)
Yes	1725 (73.7)	1548 (48.2)		1230 (73.4)	1070 (47.9)	
No	614 (26.3)	1664 (51.8)		446 (26.6)	1166 (52.1)	
Livestock			191 (3.4)			150 (3.8)
Yes	1642 (70.7)	1327 (43.0)		1157 (70.6)	893 (41.7)	
No	682 (29.3)	1757 (57.0)		482 (29.4)	1251 (58.3)	
Mine dust			191 (3.4)			150 (3.8)
Yes	123 (5.3)	190 (6.2)		85 (5.2)	129 (6.0)	
No	2201 (94.7)	2894 (93.8)		1554 (94.8)	2015 (94.0)	
Molds			191 (3.4)			150 (3.8)
Yes	1095 (47.1)	841 (27.3)		795 (48.5)	625 (29.2)	
No	1229 (52.9)	2243 (72.7)		844 (51.5)	1519 (70.8)	
Oil/Gas well fumes			191 (3.4)			150 (3.8)

Yes	543 (23.4)	689 (22.3)		387 (23.6)	467 (21.8)	
No	1781 (76.6)	2395 (77.7)		1252 (76.4)	1677 (78.2)	
Radiation			191 (3.4)			150 (3.8)
Yes	196 (8.4)	263 (8.5)		142 (8.7)	198 (9.2)	
No	2128 (91.6)	2821 (91.5)		1497 (91.3)	1946 (90.8)	
Stubble smoke			191 (3.4)			150 (3.8)
Yes	1250 (53.8)	1113 (36.1)		881 (53.8)	775 (36.1)	
No	1074 (46.2)	1971 (63.9)		758 (46.2)	1369 (63.9)	
Solvent fumes			191 (3.4)			150 (3.8)
Yes	910 (39.2)	974 (31.6)		668 (40.8)	722 (33.7)	
No	1414 (60.8)	2110 (68.4)		971 (59.2)	1422 (66.3)	
Welding fumes			191 (3.4)			150 (3.8)
Yes	1230 (52.9)	988 (32.0)		874 (53.3)	708 (33.0)	
No	1094 (47.1)	2096 (68.0)		765 (46.7)	1436 (67.0)	
Wood dust			191 (3.4)			150 (3.8)
Yes	1071 (46.1)	1078 (35.0)		752 (45.9)	801 (37.4)	
No	1253 (53.9)	2006 (65.0)		887 (54.1)	1343 (62.6)	
COVARIATES						
Age (yrs.)			48 (0.9)			21 (0.5)
50-59	1044 (44.6)	1078 (33.6)		631 (37.6)	660 (29.5)	
60-69	767 (32.8)	975 (30.4)		590 (35.2)	689 (30.8)	
70+	528 (22.6)	1159 (36.1)		455 (27.1)	887 (39.7)	
BMI (kg/m^2)			332 (5.9)			248 (6.3)
Normal (0-<25)	610 (27.4)	841 (27.7)		453 (28.7)	619 (29.4)	
Overweight (25-30)	961 (43.1)	1304 (42.9)		656 (41.5)	890 (42.3)	
Obese (>30)	659 (29.6)	892 (29.4)		470 (29.8)	597 (28.3)	
Education			130 (2.3)			80 (2.0)
≤ Grade 12	1569 (67.9)	2059 (65.2)		1051 (63.7)	1330 (60.3)	
> Grade 12	742 (32.1)	1099 (34.8)		598 (36.2)	874 (39.7)	
Marital status			78 (1.4)			33 (0.8)
Married/common law/living together	2070 (88.8)	2481 (77.8)		1504 (90.1)	1772 (79.5)	
Widowed/Divorced/separated/single/never married	261 (11.2)	709 (22.2)		166 (9.9)	458 (20.5)	

Sex			48 (0.9)			22 (0.6)
Male	1256 (53.7)	1542 (48.0)		884 (52.7)	1055 (47.2)	
Female	1083 (46.3)	1670 (52.0)		792 (47.3)	1180 (52.8)	

[#] Due to missing observations, row (variable) totals may not sum to column totals

5.3.1 Population characteristics by quadrant

Tables 5.2 and 5.3 provide the distribution selected population characteristics stratified according to the four quadrants of Saskatchewan at baseline and follow-up respectively. With the exception of the South-western corner of the province, the percentage of study participants from the remaining quadrants were higher in rural municipalities (RM) as compared to small towns at baseline (Table 5.2). Gender proportions in all four quadrants appear not overly skewed or biased towards a particular quadrant (Table 5.2). While SW (31.4%) and NE (31.5%) had the highest percentage of study participants aged 70 or older, SE (29.5%) and NW (29.3%) had the least. It is observed that the eastern corner of the province contributed the highest percentage [SE (39.8%)] of the youngest study participants as well as the least [NE (36.7%)] (Table 5.2). In all four quadrants, the majority (> 50%) of study participants were non-farm residents as compared to farm residents (Table 5.2). Table 5.3 provides an analogous description/analysis of the scenario at follow-up.

Table 5. 2 Population characteristics by quadrant at baseline

Characteristics#	**Quadrant***				
	South West	**South East**	**North East**	**North West**	**P-value**
	N = 1034	N = 1175	N = 1735	N = 1651	
Municipality, n (%)					
RM	477 (46.1)	601 (51.1)	1188 (68.5)	853 (51.7)	**<0.001****
Town	557 (53.9)	574 (48.9)	547 (31.5)	798 (48.3)	
Gender, n (%)					
Male	107 (48.0)	170 (53.5)	133 (51.6)	185 (49.2)	**0.504**
Female	116 (52.0)	148 (46.5)	125 (48.4)	191 (50.8)	
Age-groups, n (%)					
50-59	401 (38.8)	468 (39.8)	636 (36.7)	629 (38.1)	**<0.001****
60-69	308 (29.8)	360 (30.6)	553 (31.9)	539 (32.6)	
70+	325 (31.4)	347 (29.5)	546 (31.5)	483 (29.3)	
Location of home, n (%)					
Farm	372 (36.2)	439 (37.9)	856 (49.7)	672 (41.1)	**<0.001****
Non-Farm	656 (63.8)	720 (62.1)	867 (50.3)	965 (58.9)	

# Due to missing observations, variable totals may not sum to 5955

** *p*<0.05

Table 5. 3 Population characteristics by quadrant at follow-up

Characteristics[#]	Quadrant[*]				P-value
	South West	South East	North East	North West	
	N = 710	N = 799	N = 1198	N = 1149	
Municipality, n (%)					
RM	341 (48.0)	435 (54.4)	845 (70.5)	606 (52.7)	<0.001**
Town	369 (52.0)	364 (45.6)	353 (29.5)	543 (47.3)	
Gender, n (%)					
Male	355 (50.0)	392 (49.1)	598 (49.9)	575 (50.0)	0.610
Female	355 (50.0)	407 (50.9)	600 (50.1)	574 (50.0)	
Age-groups, n (%)					
50-59	256 (36.1)	281 (35.2)	354 (29.5)	373 (32.5)	<0.001**
60-69	212 (29.9)	267 (33.4)	377 (31.5)	412 (35.9)	
70+	242 (34.1)	251 (31.4)	467 (39.0)	364 (31.7)	
Location of home, n (%)					
Farm	257 (36.5)	317 (40.0)	587 (49.3)	485 (42.3)	<0.001**
Non-Farm	448 (63.5)	476 (60.0)	603 (50.7)	662 (57.7)	

[#] Due to missing observations, variable totals may not sum to 5955

** *p*<0.05

5.3.2 Distribution of colorectal cancer (CRC) by residence status

The presence or absence of CRC was determined from the baseline and follow-up questionnaires respectively. From-here-on-in, we will define cases to be individuals who self-reported a previous diagnosis of CRC by a PCG or a doctor and non-cases as individuals who self-reported otherwise. On average, cases were older than non-cases (Data not shown). The mean (± standard deviation) age of cases at baseline was 69 (± 9.8) years as compared to that of non-cases of 64 (± 10.4) years. Similarly, the mean age of cases at follow-up was 71 (± 10.2) years and that for non-cases was 65 (± 10.3) (Data not shown). Table 5.4 provides the number of study participants (%) stratified by CRC status and wave (i.e. baseline and follow-up). Overall, there was an increase in the number of CRC cases from baseline to follow-up. The percentage of CRC cases showed an increase from 29.5% at baseline to 31.8% at follow-up. However, the percentage of non-CRC cases slightly declined from 70.5% at baseline to 68.2% at follow-up.

Table 5. 4 Location of home by CRC cases and non-cases at baseline

	Baseline (2010)			Follow-up (2014)		
Location of home	**Cases (%)***	**Non-Cases (%)**	**Total (%)**	**Cases (%)**	**Non-Cases (%)**	**Total (%)**
Farm	18 (29.5)	2321 (42.3)	2339 (42.1)	21 (31.8)	1655 (43.0)	1676 (42.8)
Non-Farm	43 (70.5)	3169 (57.7)	3212 (57.9)	45 (68.2)	2191 (57.0)	2236 (57.2)
Total	61	5490	5551	66	3846	3912

* One CRC case could not be identified by the location of home (farm or non-farm);

5.3.3 Baseline and follow-up characteristics of participants by CRC status

The baseline and follow-up characteristics of study participants stratified by CRC status are presented in Table 5.5. Stratifying individuals diagnosed with CRC by age groups revealed that CRC was more common in older age groups. At baseline, the majority (1.8%) of the people diagnosed with CRC were 70 years of age or older, followed by persons in their 60's (1.2%), and the least number of CRC cases occurred in the youngest age category (0.4%). A similar trend of increasing prevalence of CRC among older age groups was observed at follow-up; 50 – 59 (0.5%), 60 – 69 (1.6%), and 70+ (2.9%). Obese individuals have the highest percentage of CRC cases both at baseline and at follow-up (Table 5.5). Individuals with the highest education being Grade 12 had the highest percentage of CRC cases. By stratifying CRC cases by marital status, the percentage of CRC cases was lower among married/common-law or cohabiting individuals (1.0%) as compared to widowed/divorced or single individuals (1.8%). This observation lasted at follow-up. A slightly higher percentage of females (1.3%) were diagnosed with CRC than males (1.0%).

Of the total number of individuals who answered "*yes*" to having being diagnosed with CRC previously in 2010, 0.8% were farm residents while 1.3% were non-farm residents. About 1.1% of CRC cases either ever lived on a farm or lived on a farm in their first year of life at baseline (Table 5.5). A higher percentage of CRC cases had mildew odor or musty smell in their homes (1.7%) as compared to those who did not at baseline (1.0%). This trend reversed in 2014 (Table 5.5). When CRC cases were stratified according to smoking status at baseline, 0.9% of them were current smokers, 1.3% of them were ex-smokers, while 1% of them never smoked. About 0.6% of the total number of CRC cases were daily consumers of alcohol at baseline while 0.9% of them never consumed alcohol (Table 5.5).

Personal and family history of cancer among CRC cases were also studied at baseline. It was realized that all 62 CRC patients had a previous personal history of cancer. About 1.2% of them had fathers with a history of cancer and while and 1.8% had mothers with a previous history of cancer. About 1.4% of individuals diagnosed with CRC had siblings with a previous history of cancer.

Since occupational exposure was of primary concern in this thesis, it was studied among individuals diagnosed with CRC. By stratifying CRC cases according to occupational exposure, it was realized that 1.0% of them were exposed to the following at work; wood dust, livestock, diesel fumes, fungicides, and grain dust respectively at baseline. A slightly lower percentage of CRC cases were exposed to oil and gas well fumes (0.7%), asbestos dust (0.8%), solvent fumes (0.9%) at work (at baseline). Approximately, 1.1% of CRC cases were exposed to molds and stubble smoke respectively, 1.2% of them were exposed to pesticides and welding fumes (respectively) during the same period. A high percentage of CRC cases were exposed to mine dust (1.3%), with an even higher percentage of them being exposed to radiation (2.2%) at their workplaces (at baseline).

Details of the above population characteristics at follow-up are described in the last column of Table 5.5 and describes the percentages of CRC cases associated with the various population characteristics. Table 5.6 present the characteristics for cases and non-cases at both baseline and follow-up respectively.

Table 5. 5 Population characteristics stratified by CRC status and survey

	Baseline (2010)		Follow-up (2014)	
Description	**Ever diagnosed with CRC**		**Ever diagnosed with CRC**	
	Yes/Total	**%**	**Yes/Total**	**%**
CONTEXTUAL FACTORS				
Socioeconomic				
Some money	42/3018	1.4	31/2392	1.3
Just enough money	9/1011	0.9	14/671	2.1
Not enough money	7/938	0.7	8/423	1.9
Quadrant				
South West	8/1034	0.8	6/710	0.8
South East	15/1175	1.3	18/799	2.3
North East	23/1735	1.3	24/1198	2.0
North West	16/1651	1.0	16/1149	1.4
Environmental:				
Household smoking				
Yes	6/791	0.8	5/395	1.3
No	55/4771	1.2	61/3538	1.7
Location of home				
Farm	18/2339	0.8	21/1676	1.3
Non-farm	43/3212	1.3	45/2236	2.0
Household structure:				
Water source				
Bottled Water	16/1501	1.1	18/1075	1.7
Deep well water (more than 100ft)	11/1391	0.8	13/919	1.4
Shallow well water (less than 100ft)	7/1134	0.6	8/747	1.1
Other sources	28/1570	1.8	27/1126	2.4
Fuel source – Natural gas				
Yes	51/3763	1.4	49/2637	1.9
No	11/1815	1.1	16/1277	1.3
Mildew odor or musty smell in home				
Yes	13/785	1.7	6/619	1.0
No	48/4628	1.0	57/3209	1.8
INDIVIDUAL FACTORS				
Smoking Status				
Current Smoker	5/571	0.9	3/344	0.9
Ex-smoker	30/2279	1.3	33/1617	2.0
Never smoker	26/2710	1.0	30/1937	1.5
Alcohol consumption				
Never	10/1166	0.9	14/738	1.9
Less than once a month	19/1208	1.6	17/804	2.1
At most 2-3 times a month	15/1355	1.1	17/938	1.8
At most 2-3 times a week	14/1174	1.2	11/881	1.2
Everyday	4/668	0.6	6/545	1.1

Physical activity				
Yes	33/3084	1.1	34/2081	1.6
No	26/2313	1.1	31/1756	1.8
Diabetes				
Yes	10/663	1.5	7/391	1.8
No	52/4846	1.1	57/3432	1.7
Early life-exposures:				
Ever-lived on a farm				
Yes	54/4875	1.1	60/3481	1.7
No	8/705	1.1	6/448	1.3
Lived on a farm in first year of life				
Yes	46/4095	1.1	55/2942	1.9
No	16/1432	1.1	11/959	1.1
Hereditary:				
Familial History of cancer:				
Personal history of cancer				
Yes	62/611	10.1	37/357	10.4
No	0	-	27/3465	0.8
Father ever had cancer				
Yes	19/1652	1.2	22/1172	1.9
No	35/3395	1.0	40/2383	1.7
Mother ever had cancer				
Yes	26/1448	1.8	24/1040	2.3
No	34/3691	0.9	38/2587	1.5
Sibling(s) ever had cancer				
Yes	21/1451	1.4	22/916	2.4
No	32/3208	1.0	32/2387	1.3
Occupational Exposures:				
At work, ever exposed to:				
Asbestos dust				
Yes	3/375	0.8	3/255	1.2
No	56/5079	1.1	61/3549	1.7
Diesel fumes				
Yes	33/3266	1.0	37/2304	1.6
No	26/2188	1.2	27/1500	1.8
Fungicides (to treat grain)				
Yes	19/1951	1.0	25/1387	1.8
No	40/3503	1.1	39/2417	1.6
Grain dust				
Yes	37/3846	1.0	39/2676	1.5
No	22/1608	1.4	25/1128	2.2
Pesticides (to kill plants and insects)				
Yes	38/3293	1.2	38/2308	1.6
No	24/2306	1.0	28/1625	1.7
Livestock				
Yes	31/2994	1.0	36/2057	1.8
No	28/2460	1.1	28/1747	1.6

Mine dust				
Yes	4/315	1.3	3/214	1.4
No	55/5139	1.1	61/3590	1.7
Molds				
Yes	21/1952	1.1	25/1427	1.8
No	38/3502	1.1	39/2377	1.6
Oil/Gas well fumes				
Yes	9/1243	0.7	13/858	1.5
No	50/4211	1.2	51/2946	1.7
Radiation				
Yes	10/465	2.2	10/343	2.9
No	49/4989	1.0	54/3461	1.6
Stubble smoke				
Yes	27/2383	1.1	31/1664	1.9
No	32/3071	1.0	33/2140	1.5
Solvent fumes				
Yes	18/1897	0.9	24/1397	1.7
No	41/3557	1.2	40/2407	1.7
Welding fumes				
Yes	26/2234	1.2	26/1589	1.6
No	33/3220	1.0	38/2215	1.7
Wood dust				
Yes	22/2157	1.0	25/1556	1.6
No	37/3297	1.1	39/2248	1.7
Covariates				
Age (yrs.)				
50-59	9/2134	0.4	7/1296	0.5
60-69	22/1762	1.2	20/1287	1.6
70+	20/1116	1.8	39/1350	2.9
BMI (kg/m^2)				
Normal (0-<25)	16/1465	1.1	20/1078	1.9
Overweight (25-30)	26/2284	1.1	19/1554	1.2
Obese (>30)	19/1560	1.2	23/1074	2.1
Education				
≤ Grade 12	43/3656	1.2	46/2395	1.9
> Grade 12	16/1857	0.9	19/1478	1.3
Marital status				
Married/common law/living together	44/4586	1.0	47/3294	1.4
Widowed/Divorced/separated/never married	18/981	1.8	19/627	3.0
Sex				
Male	27/2824	1.0	33/1950	1.7
Female	35/2774	1.3	33/1982	1.7

Table 5. 6 Population Characteristics for CRC Cases and Non-cases at Baseline and Follow-up

	Baseline (2010)		Follow-up (2014)	
Description	**CRC**		**CRC**	
	Cases (%)	**Non-cases (%)**	**Cases (%)**	**Non-cases (%)**
CONTEXTUAL FACTORS				
Socioeconomic				
Some money	42 (72.4)	2976 (60.6)	31 (58.5)	2361 (68.8)
Just enough money	9 (15.5)	1002 (20.4)	14 (26.4)	657 (19.1)
Not enough money	7 (12.1)	931 (19.0)	8 (15.1)	415 (12.1)
Quadrant				
South West	8 (12.9)	1026 (18.5)	6 (9.4)	704 (18.6)
South East	15 (24.2)	1160 (21.0)	18 (28.1)	781 (20.6)
North East	23 (37.1)	1712 (30.9)	24 (37.5)	1174 (31.0)
North West	16 (25.8)	1635 (29.5)	16 (25.0)	1133 (29.9)
Environmental:				
Location of home				
Farm	18 (29.5)	2321 (42.3)	21 (31.8)	1655 (43.0)
Non-farm	43 (70.5)	3169 (57.7)	45 (68.2)	2191 (57.0)
INDIVIDUAL FACTORS				
Smoking Status				
Current Smoker	26 (42.6)	566 (10.3)	3 (4.5)	341 (8.9)
Ex-smoker	30 (49.2)	2249 (40.9)	33 (50.0)	1584 (41.3)
Never smoker	5 (8.2)	2684 (48.8)	30 (45.5)	1907 (49.8)
Early life-exposures:				
Ever-lived on a farm				
Yes	8 (12.9)	697 (12.6)	60 (90.9)	3421 (88.6)
No	54 (87.1)	4821 (87.4)	6 (9.1)	442 (11.4)
Lived on a farm in first year of life				
Yes	46 (74.2)	4049 (74.1)	55 (83.3)	2887 (75.3)
No	16 (25.8)	1416 (25.9)	11 (16.7)	948 (24.7)
Familial History of cancer: **Ever had cancer**				
Personal				
Yes	62 (100)	549 (10.1)	37 (57.8)	320 (8.5)
No	0 (0.0)	4904 (89.9)	27 (42.2)	3438 (91.5)
Father				
Yes	19 (35.2)	1633 (32.7)	22 (35.5)	1150 (32.9)
No	35 (64.8)	3360 (67.3)	40 (64.5)	2343 (67.1)
Mother				
Yes	26 (43.3)	1422 (28.0)	24 (38.7)	1016 (28.5)
No	34 (56.7)	3657 (72.0)	38 (61.3)	2549 (71.5)
Sibling(s)				
Yes	21 (39.6)	1430 (31.0)	22 (40.7)	894 (27.5)
No	32 (60.4)	3176 (69.0)	32 (59.3)	2355 (72.5)
Occupational Exposures:				
At work, ever exposed to:				
Asbestos dust				

	Yes	3 (5.1)	372 (6.9)	3 (4.7)	252 (6.7)
	No	56 (94.9)	5023 (93.1)	61 (95.3)	3488 (93.3)
Diesel fumes					
	Yes	33 (55.9)	3233 (59.9)	37 (57.8)	2267 (60.6)
	No	26 (44.1)	2162 (40.1)	27 (42.2)	1473 (39.4)
Fungicides (to treat grain)					
	Yes	19 (32.2)	1932 (35.8)	25 (39.1)	1362 (36.4)
	No	40 (67.8)	3463 (64.2)	39 (60.9)	2378 (63.6)
Grain dust					
	Yes	37 (62.7)	3809 (70.6)	39 (60.9)	2637 (70.5)
	No	22 (37.3)	1586 (29.4)	25 (39.1)	1103 (29.5)
Pesticides					
	Yes	38 (61.3)	3255 (58.8)	38 (57.6)	2270 (58.7)
	No	24 (38.7)	2282 (41.2)	28 (42.4)	1597 (41.3)
Livestock					
	Yes	31 (52.5)	2963 (54.9)	36 (56.3)	2021 (54.0)
	No	28 (47.5)	2432 (45.1)	28 (43.8)	1719 (46.0)
Mine dust					
	Yes	4 (6.8)	311 (5.8)	3 (4.7)	211 (5.6)
	No	55 (93.2)	5084 (94.2)	61 (95.3)	3529 (94.4)
Molds					
	Yes	21 (35.6)	1931 (35.8)	25 (39.1)	1402 (37.5)
	No	38 (64.4)	3464 (64.2)	39 (60.9)	2338 (62.5)
Oil/Gas well fumes					
	Yes	9 (15.3)	1234 (22.9)	13 (20.3)	845 (22.6)
	No	50 (84.7)	4161 (77.1)	51 (79.7)	2895 (77.4)
Radiation					
	Yes	10 (16.9)	455 (8.4)	10 (15.6)	333 (8.9)
	No	49 (83.1)	4940 (91.6)	54 (84.4)	3407 (91.1)
Stubble smoke					
	Yes	27 (45.8)	2356 (43.7)	31 (48.4)	1633 (43.7)
	No	32 (54.2)	3039 (56.3)	33 (51.6)	2107 (56.3)
Solvent fumes					
	Yes	18 (30.5)	1879 (34.8)	24 (37.5)	1373 (36.7)
	No	41 (69.5)	3516 (65.2)	40 (62.5)	2367 (63.3)
Welding fumes					
	Yes	26 (44.1)	2208 (40.9)	26 (40.6)	1563 (41.8)
	No	33 (55.9)	3187 (59.1)	38 (59.4)	2177 (58.2)
Wood dust					
	Yes	22 (37.3)	2135 (39.6)	25 (39.1)	1531 (50.9)
	No	37 (62.7)	3260 (60.4)	39 (60.9)	2209 (59.1)
Covariates					
Age (yrs.)					
50-59		9 (14.5)	2125 (38.4)	7 (10.6)	1289 (33.3)
60-69		22 (35.5)	1740 (31.4)	20 (30.3)	1267 (32.8)
70+		31 (50.0)	1672 (30.2)	39 (59.1)	1311 (33.9)
BMI (kg/m^2)					

	Normal (0-<25)	16 (26.2)	1449 (27.6)	20 (32.3)	1058 (29.0)
	Overweight (25-30)	26 (42.6)	2258 (43.0)	19 (30.6)	1535 (42.1)
	Obese (>30)	19 (31.1)	1541 (29.4)	23 (37.1)	1051 (28.8)
Sex					
	Male	27 (43.5)	2797 (50.5)	33 (50.0)	1917 (49.6)
	Female	35 (56.5)	2739 (49.5)	33 (50.0)	1949 (50.4)

5.4 Research Question One (1): Prevalence Analysis

In this section, I provide results for the first research question of this thesis, which is;

"What is the prevalence of self-reported doctor-diagnosed CRC and associated risk factors in rural Saskatchewan using GEE and robust variance estimation techniques?"

5.4.1 Crude Prevalence Rate Calculation

Using equation (*) stated in section 4.5.1.1, the crude prevalence of CRC was computed. The overall crude prevalence of CRC was 1.1% at baseline and 1.7% at follow-up. When stratified by farm and non-farm residence, it was observed that the crude prevalence of CRC was almost two times lower among farm residents than among non-farm residents (0.8% vs. 1.3%) at baseline (Table 5.7). Among farmers who self-reported doctor-diagnosed CRC, the highest prevalence occurred in SE (1.4%) while the least occurred in SW (0.3) and NW (0.3) respectively (Table 5.7). Among non-farm residents, CRC prevalence was highest in NW (1.5%) and least in SW (1.1%) (Table 5.7). Overall, CRC was more prevalent in the eastern part of the quadrant [SE (1.3%) and NE (1.3%)] as compared to the western part [(SW (0.8%) and NW (1.0%))] at baseline.

Table 5. 7 Prevalence of Colorectal Cancer Stratified by Geographic Location and Farm/Non-Farm Residence at Time-point 1 (Baseline).

	Farm Dwellers		**Non-farm Dwellers**			**Total**	
Quadrant of Saskatchewan	Doctor-diagnosed self-reported Colorectal cancer		Doctor-diagnosed self-reported Colorectal cancer		Total number of missing Observations	Ever-diagnosed Colorectal Cancer	
	Yes/Total	(%)	Yes/Total	(%)	n	Yes/Total	(%)
South West	1/372	0.3	7/656	1.1	6	8/1034	0.8
South East	6/439	1.4	9/720	1.3	16	15/1175	1.3
North East	9/856	1.1	13/867	1.3	12	23/1735	1.3
North West	2/672	0.3	14/965	1.5	14	16/1651	1.0
Not identified	0/0	-	0/4	0	4	0/4	0
Total	18/2339	0.8	43/3212	1.3		62/5599	1.1

Due to missing observations in variables, totals may not sum to 5599

Table 5.8 provides the crude prevalence estimates of CRC stratified by quadrants at follow-up. Overall, SE (2.3%) reported the highest prevalence of CRC, followed by NE (2.0%) while SW (0.8%) reported the lowest. In particular, the highest prevalence of CRC occurred in NW and SE among farm (2.2%) and non-farm (3.2%) residents respectively (Table 5.8). It was also observed that CRC prevalence was lowest in SW among both farm and non-farm rural residents (0.4% and 0.8%) respectively.

Table 5. 8 Prevalence of Colorectal Cancer Stratified by Geographic Location and Farm/Non-Farm Residence at Time-point 2 (Follow-up).

	Farm Dwellers		**Non-farm Dwellers**			**Total**	
Quadrant of Saskatchewan	Doctor-diagnosed self-reported Colorectal cancer		Doctor-diagnosed self-reported Colorectal cancer		Total number of missing Observations	Ever diagnosed Colorectal Cancer	
	Yes/Total	%	Yes/Total	%	n	Yes/Total	(%)
South West	1/257	0.4	5/453	1.1	5	6/710	0.8
South East	3/317	0.9	15/476	3.2	6	18/799	2.3
North East	13/587	2.2	11/603	1.8	8	24/1198	2.0
North West	3/485	0.6	13/662	2.0	2	16/1149	1.4
Not identified	1/30	-	1/47	2.1	77	2/77	2.6
Total*	21/1676	1.3	45/2236	2.0		66/3933	1.7

* Overall, there is a significant difference between the prevalence of CRC among farm and non-farm dwellers using the chi-squared test statistic ($p < 0.05$).

Over time, the prevalence of CRC among farm residents increased from 0.8% at baseline (Table 5.7) to 1.3% at follow-up (Table 5.8), representing a 62.5% increment in CRC prevalence. On the contrary, CRC prevalence among non-farm residents increased by 43% from 1.4% at baseline to 2.0% at follow-up.

5.4.2 Adjusted Prevalence of CRC using Marginal Modeling Approach

Adjusted prevalence of CRC was determined by fitting marginal logistic regression models (adjusted for covariates) using the generalized estimating equations (GEE) approach. The GEE approach is based on the mathematical theory of multivariate quasi-likelihood. The SAS procedure GENMOD (with a clustering option) was used to identify significant risk factors for CRC prevalence both at baseline and follow-up separately as cross-sectional studies by accounting the clustering effect of individuals within a household. The GENMOD procedure was also used to determine the significant risk factors for the longitudinal changes in CRC prevalence by accounting for two layers of complexity: first, the clustering effects of more than one individual in a household; second, the within-subject correlation due to repeated measures over time. The following sections provide the results of these methods.

5.4.2.1 Marginal Modeling Approach for Cross-sectional Data

Baseline and follow-up survey data were separately analyzed as cross-sectional data to identify risk factors for CRC at the two time-points. As indicated in section 5.3.3, all the CRC cases had a previous personal history of cancer at baseline resulting in a frequency of zero (0) for the cell representing CRC cases without a previous history of cancer in a crosstab analysis. This resulted in a quasi-complete separation problem in the model fitting process. To avoid this, a decision was made to exclude participants' personal history of cancer in all the regression models contained in this thesis. Table 5.9 provides the univariable associations between baseline and follow-up characteristics and CRC risk.

CRC cancer risk was assessed at baseline. A dose-response relationship was observed for age. Individuals aged 60 – 69 years were almost 3 times more likely to develop CRC as compared to those in their 50's. Adults who were 70 years or older were 4.38 times more at risk of developing CRC than adults aged 50 – 59 years. Marital status was protective for CRC (Table 5.9). Individuals exposed to radiation at work were 2.22 times more likely to develop CRC as compared to non-exposed individuals. Oil/Gas well fumes (OR = 0.61) and grain dust (0.70) were associated with a lower CRC risk (Table 5.9). Individuals whose mothers had a previous history of cancer were about 2 times more likely to be diagnosed with CRC as compared to individuals whose mothers had no such history (OR = 1.97). A similarly elevated risk was observed among individuals with a sibling(s) who had a previous history of cancer when compared with individuals with siblings having no such history (OR = 1.46). The use of natural gas as a household source of fuel more than doubled the risk of developing CRC (OR = 2.25). Increased risk of developing CRC was observed among individuals whose homes had mildew odor or musty smell as compared to others whose homes did not have such odor (OR = 1.61).

Households' source of water supply was found to significantly influence the risk of CRC. Among households whose source of water was deep and shallow wells (as compared with bottled water), a reduced risk of CRC was observed (OR = 0.74 and 0.58 respectively). However, a higher risk was found among individuals whose households water supply was from lakes, dug-outs, rivers, etc. (OR = 1.94). Individuals with "*just enough money*" left at the end of the month were 0.64 times less likely to be diagnosed with CRC as compared to those with "*some money*". Among individuals with "not enough money", a similarly low risk of CRC was observed (OR = 0.53), although not statistically significant.

A similar univariable analysis was done at follow-up. Like at baseline, a dose-response relationship between CRC risk and age was found (Table 5.9). Individuals in their 60s were 2.91 times more likely to develop CRC as compared to those in their 50s. When compared with individuals aged 50 – 59 years, the risk of developing CRC among 70 year- or older adults was more than five-fold (OR = 5.48). Obese individuals were slightly more likely to develop CRC as compared to normal-weighted individuals (OR = 1.16). Married or common-law partners had a reduced risk of developing CRC as compared to widowed/divorced/separated or individuals who never married (OR = 0.46). Individuals with the highest education attained being Grade 12 were more likely to be diagnosed with CRC as compared to those with higher education. While exposure to radiation increased CRC risk, grain dust was found to lower CRC risk (OR = 1.90 vs. OR = 0.65). Having a mother or sibling with a history of cancer increased the risk of developing CRC cancer as compared to having a mother and/or sibling with no such history (OR = 1.59 and OR = 1.81 respectively). Living on a farm in one's first year of life was associated with a higher risk of developing CRC (OR = 1.64). In households where natural gas was the source of fuel, an increased CRC risk was observed (OR = 14.9). Quadrant was found to increase the risk of CRC. For example; residents of SE, NE, and NW were 2.70, 2.40, and 1.66 times (respectively) more likely to develop CRC as compared to residents of SW (Table 5.9).

Table 5. 9 Univariable Analysis of the Dependency of Colorectal Cancer on Contextual Factors, Individual Factors, and Covariates by Odds Ratio (OR), 95% CI, and P-value by Time-point

Predictor	**Baseline (2010)**		**Follow-up (2014)**	
Description	**Unadjusted Odds Ratio (OR)* (95% CI)**	**P-value#**	**Unadjusted Odds Ratio (OR)* (95% CI)**	**P-value#**
CONTEXTUAL FACTORS				
Socioeconomic		0.19		0.27
Some money	1.00		1.00	
Just enough money	0.64 (0.31, 1.31)		1.62 (0.86, 3.06)	
Not enough money	0.53 (0.24, 1.19)		1.47 (0.67, 3.21)	
Quadrant		0.50		0.13
South West	1.00		1.00	
South East	1.66 (0.70, 3.94)		2.70 (1.07, 6.83)	
North East	1.72 (0.77, 3.86)		2.40 (0.98, 5.88)	
North West	1.25 (0.54, 2.94)		1.66 (0.67, 4.24)	
Environmental:				
Household smoking		0.33		0.50
Yes	0.66 (0.28, 1.53)		0.73 (0.29, 1.83)	
No	1.00		1.00	
Location of home		0.05		0.07
Farm	0.57 (0.33, 0.99)		0.62 (0.37, 1.04)	
Non-farm	1.00			
Household structure:				
Water source				0.15
Bottled Water	1.00	0.02	1.00	
Deep well water (more than 100ft)	0.74 (0.34, 1.60)		0.84 (0.41, 1.72)	

Shallow well water (less than 100ft)	0.58 (0.24, 1.40)		0.64 (0.28, 1.47)	
Other sources	1.69 (0.91, 3.12)		1.44 (0.79, 2.63)	
Fuel source – Natural gas		0.02		0.17
Yes	2.25 (1.17, 4.33)		1.49 (0.85, 2.63)	
No	1.00			
Mildew odor or musty smell in home		0.13		0.15
Yes	1.61 (0.87, 2.97)		0.54 (0.23, 1.26)	
No	1.00		1.00	
INDIVIDUAL FACTORS				
Smoking Status		0.42		0.25
Current Smoker	0.91 (0.35, 2.38)		0.56 (0.17, 1.84)	
Ex-smoker	1.38 (0.81, 2.33)		1.32 (0.81, 2.18)	
Never smoker	1.00		1.00	
Alcohol consumption		0.34		0.51
Never	1.00		1.00	
Less than once a month	1.85 (0.86, 3.99)		1.12 (0.55, 2.28)	
At most 2-3 times a month	1.30 (0.58, 2.89)		0.95 (0.47, 1.95)	
At most 2-3 times a week	1.40 (0.62, 3.15)		0.65 (0.30, 1.45)	
Everyday	0.70 (0.22, 2.23)		0.58 (0.22, 1.50)	
Physical activity		0.85		0.85
Yes	0.95 (0.56, 1.59)		1.08 (0.49, 2.38)	
No	1.00		1.00	
Diabetes		0.32		0.75
Yes	1.41 (0.71, 2.79)		0.92 (0.57, 1.51)	
No	1.00		1.00	
Early life-exposures:				
Ever-lived on a farm		0.95		0.55
Yes	0.98 (0.46, 2.06)		1.29 (0.55, 3.01)	
No	1.00			
Lived on a farm in first year of life		0.99		0.14
Yes	1.01 (0.57, 1.78)		1.64 (0.86, 3.15)	
No	1.00			
Hereditary				
Familial History of cancer:				
Father ever had cancer		0.70		0.67
Yes	1.12 (0.64, 1.96)		1.12 (0.66, 1.89)	
No	1.00		1.00	
Mother ever had cancer		0.01		0.08
Yes	1.97 (1.18, 3.28)		1.59 (0.95, 2.65)	
No	1.00			
Sibling(s) ever had cancer		0.18		0.03
Yes	1.46 (0.84, 2.54)		1.81 (1.05, 3.13)	
No	1.00		1.00	
Occupational Exposures:				
At work, ever exposed to:				
Asbestos dust	0.72 (0.23, 2.32)	0.59	0.68 (0.21, 2.18)	0.52
Diesel fumes	0.85 (0.51, 1.42)	0.53	0.89 (0.54, 1.47)	0.65
Fungicides (to treat grain)	0.85 (0.49, 1.47)	0.57	1.12 (0.68, 1.86)	0.66

Grain dust	0.70 (0.41, 1.19)	0.19	0.65 (0.39, 1.08)	0.10
Pesticides (to kill plants and insects)	1.11 (0.66, 1.86)	0.69	0.96 (0.58, 1.56)	0.85
Livestock	0.91 (0.54, 1.52)	0.72	1.09 (0.66, 1.80)	0.73
Mine dust	1.19 (0.43, 3.30)	0.74	0.82 (0.26, 2.65)	0.74
Molds	0.99 (0.58, 1.69)	0.96	1.07 (0.64, 1.77)	0.80
Oil/Gas well fumes	0.61 (0.30, 1.24)	0.17	0.87 (0.47, 1.61)	0.67
Radiation	2.22 (1.11, 4.40)	0.02	1.90 (0.95, 3.76)	0.07
Stubble smoke	1.09 (0.65, 1.82)	0.75	1.21 (0.74, 1.99)	0.45
Solvent fumes	0.82 (0.47, 1.43)	0.49	1.03 (0.62, 1.72)	0.90
Welding fumes	1.14 (0.68, 1.91)	0.63	0.95 (0.58, 1.58)	0.85
Wood dust	0.91 (0.53, 1.54)	0.72	0.93 (0.56, 1.54)	0.76
Covariates				
Age (yrs.)		<0.001		<0.001
50-59	1.00		1.00	
60-69	2.99 (1.37, 6.50)		2.91 (1.23, 6.90)	
70+	4.38 (2.08, 9.21)		5.48 (2.44, 12.28)	
BMI (kg/m^2)		0.95		0.17
Normal (0-<25)	1.00		1.00	
Overweight (25-30)	1.04 (0.56, 1.95)		0.66 (0.35, 1.23)	
Obese (>30)	1.12 (0.57, 2.18)		1.16 (0.63, 2.12)	
Education		0.29		0.14
≤ Grade 12	1.37 (0.77, 2.44)		1.50 (0.88, 2.58)	
> Grade 12	1.00		1.00	
Marital status		0.02		0.01
Married/common law/living together	0.52 (0.30, 0.90)		0.46 (0.27, 0.79)	
Widowed/Divorced/separated/single/never married	1.00		1.00	
Sex		0.28		0.95
Male	1.00		1.00	
Female	1.32 (0.80, 2.20)		0.98 (0.60, 1.60)	

* Clustering within households was accounted using multi-level dichotomous logistic regression using generalized estimating equations (GEE);

# Statistical significance is determined at p<0.25. P-values of ORs significantly different from 1 are bold-faced.

Significant variables including important covariates in the univariable analysis were retained for the multivariable analysis. After adjusting for covariates, significant risk factors for the prevalence of CRC risk at baseline include quadrant, farm residence, water source, smoking status, family history of cancer, grain dust, radiation, age, BMI, education, marital status, and gender. Clustering effects of more than one individual within a household were accounted for using multi-level dichotomous logistic regression based on generalized estimating equations (GEE) approach. Table 5.10 provides the results for the multivariate analysis of the dependency of colorectal cancer on contextual factors, individual factors, and covariates by adjusted odds ratio (OR), 95% CI by time-point. The last column of Table 5.10 provides the adjusted OR for significant determinants of CRC risk at follow-up.

Table 5. 10 Multivariable Analysis of the Dependency of Colorectal Cancer on Contextual Factors, Individual Factors, and Covariates by Odds Ratio (OR), 95% CI, and P-value by Time-point

	Baseline (2010)		**Follow-up (2014)**	
Predictor	$\widehat{\beta}[SE(\widehat{\beta})]$	**Adjusted Odds Ratio (OR)* (95% CI)**	$\widehat{\beta}[SE(\widehat{\beta})]$	**Adjusted Odds Ratio (OR)* (95% CI)**
CONTEXTUAL FACTORS				
Quadrant				
South West	Ref.		Ref.	-
South East	0.58 (0.53)	1.79 (0.63, 5.09)	0.54 (0.64	1.72 (0.50, 5.98)
North East	0.16 (0.54)	1.18 (0.41, 3.62)	0.69 (0.59)	1.99 (0.62, 6.36)
North West	0.52 (0.52)	1.69 (0.61, 4.64)	0.40 (0.66)	1.50 (0.41, 5.47)
Environmental:				
Location of home				
Farm	-0.45 (0.39)	0.64 (0.30, 1.34)	0.01 (0.49)	1.01 (0.39, 2.64)
Non-farm	Ref.	-	Ref.	-
Household structure:				
Water source				
Bottled Water	-	-	Ref.	-
Deep well water (more than 100ft)			-0.07 (0.53)	0.93 (0.33, 2.62)
Shallow well water (less than 100ft)			-0.14 (0.55)	0.89 (0.30, 2.52)
Other sources			0.19 (0.49)	1.21 (0.46, 3.20)
Fuel source – Natural gas				
Yes	-	-	0.30 (0.52)	1.35 (0.49, 3.74)
No			Ref.	-
Mildew odor or musty smell in home				
Yes	-	-	-0.27 (0.55)	0.76 (0.26, 2.26)
No			Ref.	-
INDIVIDUAL FACTORS				
Smoking Status				
Current Smoker	0.66 (0.54)	1.94 (0.68, 5.52)	-0.45 (0.73)	0.64 (0.15, 2.68)
Ex-smoker	0.26 (0.36)	1.30 (0.64, 2.65)	0.48 (0.38)	1.62 (0.77, 3.42)
Never smoker	Ref.		Ref.	-
Lived on a farm in first year of life				
Yes	-	-	0.34 (0.44)	1.40 (0.59, 3.30)
No			Ref.	-
Hereditary:				
Familial History of cancer:				
Father ever had cancer				

Yes	0.07 (0.33)	1.08 (0.57, 2.04)	-	-
No	Ref.			
Mother ever had cancer				
Yes	0.87 (0.33)	2.40 (1.26, 4.57)	0.14 (0.37)	1.15 (0.55, 2.40)
No	Ref.		Ref.	-
Sibling(s) ever had cancer				
Yes	-0.04 (0.35)	0.96 (0.48, 1.89)	0.51 (0.35)	1.67 (0.84, 3.28)
No	Ref.		Ref.	-
Occupational Exposures:				
At work, ever exposed to:				
Grain dust	-0.35 (0.39)	0.71 (0.33, 1.51)	-	-
Radiation	0.84 (0.44)	2.31 (0.99, 5.41)	-	-
Covariates				
Age (yrs.)				
50-59	Ref.		Ref.	-
60-69	1.20 (0.49)	3.31 (1.26, 8.66)	0.72 (0.60)	2.05 (0.63, 6.69)
70+	1.75 (0.48)	5.73 (2.22, 14.74)	0.58 (0.58)	1.78 (0.57, 5.52)
BMI (kg/m^2)				
Normal (0-<25)	Ref.		Ref.	-
Overweight (25-30)	-0.13 (0.40)	0.88 (0.40, 1.94)	-0.51 (0.50)	0.60 (0.23, 1.61)
Obese (>30)	0.21 (0.41)	1.23 (0.56, 2.74)	0.15 (0.47)	1.16 (0.46, 2.91)
Education				
≤ Grade 12	0.27 (0.37)	1.31 (0.63, 2.74)	0.01 (0.41)	1.01 (0.45, 2.23)
> Grade 12	Ref.		Ref.	-
Marital status				
Married/common law/living together	-0.36 (0.33)	0.70 (0.36, 1.36)	-0.52 (0.41)	0.60 (0.27, 1.34)
Widowed/Divorced/separated/single/never married			Ref.	-
Sex				
Male	Ref.		Ref.	-
Female	0.02 (0.35)	1.02 (0.52, 2.01)	-0.64 (0.42)	0.53 (0.23, 1.19)

* Clustering within households was accounted using multi-level univariable dichotomous logistic regression using generalized estimating equations (GEE);

# Statistical significance is determined at p<0.25. P-values of ORs significantly different from 1 are bold-faced.

At the end of the model building process, the following two final models were arrived at, one for the baseline and follow-up analysis respectively. No significant interactions and/or confounders were found.

Final Model 1: Probability of CRC at Baseline

$logit[Pr$(CRC = 1)] = 0.72 + 0.58*(SE) + 0.16*(NE) + 0.52*(NW) – 0.45*(Farm) + 0.66*(Current smoker) + 0.26*(Ex-smoker) + 0.07*(Father ever had cancer) + 0.87*(Mother ever had cancer) – 0.04*(Sibling ever had cancer) – 0.35*(Exposure to grain dust) + 0.84*(Exposure to radiation) + 1.20*(Age 60 – 69) + 1.75*(Age ≥ 70) – 0.13*(Overweight) + 0.21*(Obese) + 0.27*(≤ Grade 12) – 0.36*(Married or common-law) + 0.02*(Female gender) (Table 5.10).

Final Model 1: Probability of CRC at Follow-up

$logit[Pr$(CRC = 1)] = - 0.54 + 0.54*(SE) + 0.69*(NE) + 0.40*(NW) + 0.01*(Farm) – 0.07*(Deep well) – 0.14*(Shallow well) + 0.19*(Other source) + 0.30*(Natural Gas) – 0.27*(Mildew odor or musty smell in home) – 0.45*(Current smoker) + 0.48*(Ex-smoker) + 0.34*(Lived on farm in first year of life) + 0.14*(Mother ever had cancer) + 0.51*(Sibling ever had cancer) + 0.72*(Age 60 – 69) + 0.58*(Age ≥ 70) – 0.51*(Overweight) + 0.51*(Obese) + 0.01*(≤ Grade 12) – 0.52*(Married or common-law) – 0.64*(Female gender) (Table 5.10).

For the purposes of prediction, a decision was made to fit a model containing only variables with p<0.05 or borderline significance and variables of rural exposure importance such as location of home (farm vs non-farm and rural geographical location (quadrant). The resulting model for baseline variables is shown below in Table 5.11 and no significant interaction was found. The corresponding model for follow-up was not fitted because no variable had p<0.05 or borderline significance.

Table 5. 11 Predicting Probability of CRC at baseline

	Baseline (2010)		
Predictor	$\widehat{\beta}[SE(\widehat{\beta})]$	**P-value**	**Adjusted Odds Ratio (OR)* (95% CI)**
Intercept	-3.18 (0.44)	<0.001	0.04 (0.02, 0.10)
Quadrant			
South West	Ref.		
South East	-0.18 (0.44)	0.68	0.83 (0.35, 1.98)
North East	0.21 (0.39)	0.59	1.23 (0.58, 2.62)
North West	0.30 (0.34)	0.38	1.35 (0.69, 2.64)

Location of home			
Farm	0.40 (0.30)	0.18	1.49 (0.84, 2.65)
Non-farm	Ref.		-
Age			
50 - 59	Ref.		-
60 - 69	-1.66 (0.40)	<0.001	0.19 (0.09, 0.41)
70+	-0.66 (0.30)	0.03	0.52 (0.29, 0.93)
Gender			
Male	Ref.		-
Female	-0.18 (0.27)	0.02	0.55 (0.33, 0.94)
Exposure to Radiation			
Yes	-0.62 (0.37)	0.10	0.54 (0.26, 1.12)
No	Ref.		-

The logit form of the model contained in Table 5.11 is thus written as follows:

Final Model 1 for Prediction: Probability of CRC at Baseline

$logit[Pr($CRC = 1)] = -3.18 – 0.18*(SE) + 0.21*(NE) + 0.30*(NW) – 0.40*(Farm) – 0.59*(Mother ever had cancer) – 0.62*(Exposure to radiation) – 1.66*(Age 60 – 69) – 0.66*(Age ≥ 70) – 0.18*(Female gender).

5.4.2.2 Marginal Modeling Approach for Longitudinal Changes in CRC Prevalence

To determine the longitudinal changes in CRC prevalence, the marginal modeling approach used for analyzing cross-sectional data was extended to longitudinal data analysis. Information on baseline and follow-up characteristics were combined into a single longitudinal data file using PERSON ID as the matching variable. ORs were calculated with the help of the GENMOD procedure in SAS using the GEE approach. An exchangeable correlation matrix was specified, and the corresponding convergence criteria satisfied with the Hessian matrix being positive definite.

Standard model building procedures were used in the selection of variables for the final model. Univariable analysis was done using a dichotomous CRC outcome (*yes/no*) with putative risk factors for CRC prevalence. Variables with p<0.25 together with biological or clinical relevance were candidates for the multivariable model. Table 5.12 provides the crude and adjusted ORs associated with determinants of the longitudinal changes in CRC prevalence.

In the final multivariable model, significant determinants of CRC prevalence included quadrant, location of home (farm or non-farm), water source, use of natural gas, smoking status, having a mother or sibling with a previous history of cancer, exposure to grain dust and radiation, age, BMI, educational and marital status, and gender.

The odds of developing CRC among residents in SE, NE, and NW were 2.06, 1.72, and 1.18 (respectively) times higher as compared to residents in SW. Farm residents were at a higher risk of developing CRC as compared to non-farm residents (OR = 1.26). Deep well (OR = 0.70) and Shallow well (OR = 0.76) water were associated with a lower risk of developing CRC as compared to bottled water. However, the consumption of water from rivers, springs, dugouts, and lakes had a higher risk of developing CRC as compared to the consumption of bottled water (OR = 1.27). Households using natural gas as fuel were 1.74 times more likely to be diagnosed with CRC as compared to those who did not. Being a current smoker was associated with slightly low risk while ex-smokers had an elevated risk of being diagnosed with CRC as compared to never smokers (OR = 0.98 vs. 1.74). Individuals with mothers or siblings with a previous history of cancer were 1.62 and 1.29 (respectively) more likely to be diagnosed with CRC as compared to individuals having mothers/siblings with no such history. While exposure to radiation was associated with higher odds of developing CRC (OR = 1.79; 95% CI: 0.87, 3.69), exposure to grain dust was associated with lower odds (OR = 0.76; 95% CI: 0.42, 1.39), although these associations were not statistically significant.

A significant dose-response relationship between CRC risk and age was observed. Individuals aged 60 – 69 were about two-and-a-half times more likely to be diagnosed with CRC as compared individuals aged 50 – 59 years (OR = 2.40; 95% CI: 1.30 – 4.44). Similarly, individuals aged 70 years or older had an even higher risk of developing CRC relative to those aged 50 – 59 (OR = 3.73; 95% CI: 1.99 – 6.98). Obese individuals were 1.17 times more likely to be diagnosed with CRC as compared to normal-weighted individuals (Table 5.12). Lower education status was associated with a higher risk of developing CRC when compared with higher education status (OR = 1.05). Marital status was found to be protective against CRC risk (OR = 0.61). Females were associated with lower risk of being diagnosed with CRC as compared to males although this association was not significant (OR = 0.84) (Table 5.12).

Table 5. 12 Longitudinal Determinants of CRC Prevalence by Odds Ratio (OR), 95% CI, and P-value

	Univariable		**Multivariable**
Predictor at baseline and follow-up	**Crude Odds Ratio (OR)* (95% CI)**	**P-value**	**Adjusted Odds Ratio (OR)* (95% CI)**
CONTEXTUAL FACTORS			
Socioeconomic		0.78	
Some money	1.00		
Just enough money	0.98 (0.64, 1.51)		-
Not enough money	0.83 (0.48, 1.41)		
Quadrant		0.07	
South West	1.00		1.00
South East	2.18 (1.02, 4.66)		2.06 (0.88, 4.82)
North East	1.99 (0.96, 4.12)		1.71 (0.77, 3.81)
North West	1.30 (0.59, 2.84)		1.18 (0.48, 2.90)
Environmental:			
Household smoking		0.18	
Yes	0.66 (0.37, 1.21)		-
No	1.00		
Location of home		0.05	
Farm	0.69 (0.46, 1.03)		1.26 (0.64, 2.48)
Non-farm	1.00		1.00
Household structure:			
Water source		0.02	
Bottled Water	1.00		1.00
Deep well water (more than 100ft)	0.87 (0.49, 1.56)		0.70 (0.43, 1.14)
Shallow well water (less than 100ft)	0.68 (0.41, 1.13)		0.76 (0.41, 1.38)
Other sources	1.57 (1.02, 2.46)		1.27 (0.83, 1.94)
Fuel source – Natural gas		0.08	
Yes	1.47 (0.91, 2.39)		1.74 (0.88, 3.43)
No	1.00		1.00
Mildew odor or musty smell in home		0.30	
Yes	1.28 (0.85, 1.95)		-
No	1.00		
INDIVIDUAL FACTORS			
Smoking Status		0.03	
Current Smoker	0.62 (0.22, 1.70)		0.98 (0.40, 2.44)
Ex-smoker	1.53 (0.99, 2.37)		1.37 (0.79, 2.39)
Never smoker	1.00		1.00
Alcohol consumption		0.33	
Never	1.00		

Less than once a month	1.39 (0.82, 2.35)		-
At most 2-3 times a month	1.20 (0.66, 2.19)		
At most 2-3 times a week	0.79 (0.37, 1.67)		
Everyday	0.84 (0.42, 1.67)		
Physical activity		0.42	
Yes	0.86 (0.61, 1.23)		-
No	1.00		
Diabetes		0.60	
Yes	1.20 (0.64, 2.26)		-
No	1.00		
Early life-exposures:			
Ever-lived on a farm		0.82	
Yes	1.07 (0.60, 1.92)		-
No	1.00		
Lived on a farm in first year of life		0.29	
Yes	1.30 (0.77, 2.22)		-
No	1.00		
Hereditary:			
Familial History of cancer:			
Father ever had cancer		0.68	
Yes	1.10 (0.69, 1.77)		-
No	1.00		
Mother ever had cancer		0.03	
Yes	1.70 (1.08, 2.66)		1.62 (0.96, 2.74)
No	1.00		1.00
Sibling(s) ever had cancer		0.14	
Yes	1.48 (0.91, 2.42)		1.29 (0.76, 2.18)
No	1.00		
Occupational Exposures:			
At work, ever exposed to:			
Asbestos dust	0.69 (0.25, 1.97)	0.42	-
Diesel fumes	0.87 (0.56, 1.35)	0.53	-
Fungicides (to treat grain)	0.95 (0.61, 1.50)	0.85	-
Grain dust	0.69 (0.44, 1.08)	0.13	0.76 (0.42, 1.39)
Pesticides (to kill plants and insects)	0.73 (0.51, 1.04)	0.04	-
Livestock	0.98 (0.64, 1.52)	0.94	-
Mine dust	1.20 (0.54, 2.69)	0.68	-
Molds	0.99 (0.63, 1.57)	0.98	-
Oil/Gas well fumes	0.75 (0.44, 1.28)	0.25	-
Radiation	2.18 (1.22, 3.91)	0.05	1.79 (0.87, 3.69)
Stubble smoke	1.15 (0.75, 1.78)	0.53	-
Solvent fumes	0.93 (0.59, 1.47)	0.77	-
Welding fumes	1.04 (0.67, 1.61)	0.87	-
Wood dust	0.99 (0.63, 1.55)	0.97	-
Covariates			
Age (yrs.)		<0.001	

50-59	1.00		1.00
60-69	3.07 (1.67, 5.66)		2.40 (1.30, 4.44)
70+	4.82 (2.59, 8.99)		3.73 (1.99, 6.98)
BMI (kg/m^2)		0.27	
Normal (0-<25)	1.00		1.00
Overweight (25-30)	0.83 (0.49, 1.39)		0.86 (0.50, 1.52)
Obese (>30)	1.22 (0.66, 2.24)		1.17 (0.62, 2.18)
Education		0.21	
≤ Grade 12	1.34 (0.81, 2.24)		1.05 (0.60, 1.87)
> Grade 12	1.00		
Marital status		0.01	
Married/common law/living together	0.51 (0.32, 0.80)		0.61 (0.37, 1.01)
Widowed/Divorced/separated/single/never married	1.00		
Sex		0.66	
Male	1.00		
Female	1.10 (0.72, 1.68)		0.84 (0.49, 1.45)

* Repeated measurements for individuals was accounted for using PROC GENMOD using generalized estimating equations (GEE)

For the purposes of prediction, a decision was also made to fit another model that will include only variables with p<0.05 or borderline significance as contained in the model of Table 5.10. The resulting model is presented Table 5.13 below. No significant interactions were found.

Table 5. 13 Predicting Probability of CRC for Longitudinal Changes

Predictor	$\hat{\beta}[SE(\hat{\beta})]$	**P-value**	**Adjusted Odds Ratio (OR)* (95% CI)**
Intercept	-6.24 (0.49)	<0.001	0.002 (0.0008, 0.005)
Quadrant			
South West	Ref.		-
South East	0.75 (0.39)	0.06	2.11 (0.98, 4.51)
North East	0.65 (0.37)	0.08	1.91 (0.93, 3.92)
North West	0.06 (0.41)	0.88	1.06 (0.47, 2.38)
Location of home			
Farm	-0.15 (0.27)	0.59	0.86 (0.50, 1.47)
Non-farm	Ref.		-
Age			
50 - 59	Ref.		-
60 - 69	1.06 (0.34)	<0.001	2.90 (1.49, 5.64)
70+	1.64 (0.34)	<0.001	5.20 (2.67, 10.13)
Gender			
Male	Ref.		-

Female	0.10 (0.22)	0.67	1.10 (0.71, 1.71)
Exposure to Radiation			
Yes	0.60 (0.31)	0.06	1.83 (0.99, 3.39)
No	Ref.		-
Natural Gas			-
Yes	0.35 (0.31)	0.26	1.41 (0.77, 2.60)
No	Ref.		-

Final Model 1 for Prediction: Probability of CRC for Longitudinal Changes

$logit[Pr($CRC = 1)] = –6.24 + 0.74*(SE) + 0.65*(NE) + 0.06*(NW) – 0.15*(Farm) + 0.37*(Mother ever had cancer) + 0.60*(Exposure to radiation) + 1.06*(Age 60 – 69) + 1.65*(Age $\geq$ 70) + 0.10*(Female gender) + 0.35*(Natural gas).

5.5 Research Question Two (2): Incidence Analysis

In this section, I provide results for objective 1 of this thesis, by answering the research question;

"What is the incidence of self-reported doctor-diagnosed CRC in rural Saskatchewan using survival analysis techniques?"

5.5.1 Crude Incidence Calculation

Using equation (**) specified in Section 4.6.3.1.1, the crude incidence rate of CRC was calculated. For the incidence analysis, the focus was on self-reported doctor diagnosed new CRC cases over the four-year period (2010 – 2014). Figure 4.2 illustrates the selection of the valid sample size for the calculation of the crude incidence rate. Table 5.14 provides incident cases of CRC stratified by time-point/wave. The results show that, at the end of the follow-up survey, 27 newly diagnosed CRC cases (i.e. incident cases) were observed. There were 3,381 censored observations at this period. The crude incidence rate of CRC is therefore computed as;

$$\text{Crude incidence} = \frac{27}{3408} \text{ x } 100 = 0.8\%.$$

Table 5. 14 New CRC cases stratified by time-point/waves

Time-point/Wave	Event/New cases	Censored cases	Total
Baseline (2010)	*	3,435	3,435
Follow-up (2014)	27	3,381	3,408
Total	27	6,816	6,843

5.5.2 Adjusted Incidence Analysis using Cox's Proportional Hazards Regression Model

To determine risk factors for the incidence of CRC, the Cox proportional hazards (PH) regression model was used. The PH model used primarily to establish the most parsimonious relationship between CRC incidence and risk factors. At the univariable stage, the proportionality assumption was satisfied for all covariates and crude hazard rates were calculated. Variables with p-value <0.25 were candidates for the multivariable Cox PH regression model. Clinically relevant variables were retained in the multivariable models even when they were not significant during the univariable analysis. Standard model building techniques were used in the selection of the final multivariable model. Statistical significance was defined at p-value <0.05. Fitted models were tested using Schoenfeld residuals.

Various STATA commands were used to perform the survival analysis. The STSET was used to declare the CRC incidence data as survival-time data. The STCOX command was used to fit the proportional hazards model. Clustering effects of more than one individual in a household were accounted for using the clustered estimator command *VCE(CLUSTER HOUSEID)*, where *HOUSEID* is the within-subject variable in this study. The STPHTEST command was used to test the proportional hazards model assumption.

In this thesis, the exact time of the occurrence of colorectal cancer (CRC) was not known. However, the time interval in which the disease could have occurred is known. This becomes a discrete process and results in interval censored variables. As a result, a discrete-time survival analysis where each person is followed equally for four (4) years was conducted. The results show that quadrant, use of natural gas, living on a farm in one's first year of life, exposure to oil/gas well fumes and radiation, increasing age, increasing BMI, and female gender predict CRC incidence. Table 5.15 provides the crude and adjusted hazard ratios (HR) associated with risk factors for the incidence of CRC.

When adjusted for other covariates in the PH model, a dose-response relationship was observed between age and the time-to-CRC. The hazard ratio increased by 1.13 and 3.22 times in individuals aged 60 – 69 and 70 or older respectively as compared to people aged 50-59. The HR of CRC incidence for overweight individuals reduced marginally by 9% compared to normal weighted individuals (HR = 0.91). This reduction was not statistically significant (Table 5.15). The HR of CRC was 1.29 times higher among obese individuals as compared to normal weighted individuals. The HR of CRC among residents of SE, NE, and NW were 3.2, 4.10, and 1.71 times higher than among residents in the SW (Table 5.15).

The impact of occupational exposure on the time until the development of CRC was investigated. Exposure to oil/gas well fumes almost doubled the HR of CRC (HR = 1.81). A similar elevated HR of CRC was associated with exposure to radiation as compared (HR = 1.69) individuals who were not exposed. Living on a farm during one's first year of life increased the HR of CRC by 1.55 times relative to individuals who did not live their first year of life on a farm. The HR of CRC among households that use natural gas was 1.62 times higher than those that did not. There was a significant interaction between quadrant and gender. However, the confidence interval associated with this interaction term was very wide and also rendered the Cox PH regression coefficients highly unstable. A decision was made to exclude the interaction from the final Cox's PH multivariable model and the results not presented.

Table 5. 15 Cox PH regression analysis of CRC incidence during 4 years of follow-up according to baseline characteristics

	Univariable		Multivariable
Predictor at baseline	**Crude HR (95% CI)**	**P-value**	**Adjusted HR (95% CI)**
CONTEXTUAL FACTORS			
Socioeconomic		0.99	
Some money	Ref.		*
Just enough money	2.86 (0.84 – 13.66)		
Not enough money	3.82 (1.26 – 21.52)		
Quadrant		0.16	
South West	Ref.		Ref.
South East	3.09 (0.64 – 14.78)		3.2 (0.68 – 15.18)
North East	3.74 (0.84 – 16.45)		4.10 (0.95 – 17.72)
North West	1.52 (0.30 – 7.80)		1.71 (0.33 – 8.86)
Environmental:			
Household smoking		0.92	
Yes	0.94 (0.38 – 3.10)		*
No	Ref.		
Location of home		0.41	
Farm	0.72 (0.33 – 1.57)		*
Non-farm	Ref.		
Household structure:			
Water source		0.68	
Bottled Water	Ref.		*
Deep well water (more than 100ft)	0.95 (0.37 – 2.45)		
Shallow well water (less than 100ft)	0.57 (0.18 – 1.83)		
Other sources	0.61 (0.21 – 1.82)		
Fuel source – Natural gas		0.23	
Yes	1.75 (0.71 – 4.31)		1.62 (0.64 – 4.14)
No	Ref.		Ref.
Mildew odor or musty smell in home		0.16	
Yes	0.24 (0.03 – 1.75)		*
No	Ref.		
INDIVIDUAL FACTORS			
Smoking Status		0.85	
Current Smoker	1.33 (0.38 – 4.62)		*
Ex-smoker	0.92 (0.41 – 2.06)		
Never smoker	Ref.		
Alcohol consumption		0.29	
Never	Ref.		*
Less than once a month	0.38 (0.12 – 1.23)		
At most 2-3 times a month	0.40 (0.14 – 1.17)		
At most 2-3 times a week	0.63 (0.24 – 1.69)		

Everyday	0.32 (0.07 – 1.49)		
Physical activity		0.33	*
Yes	0.68 (0.32 – 1.47)		
No	Ref.		
Diabetes		0.57	
Yes	0.67 (0.18 – 2.80)		*
No	Ref.		
Early life-exposures:			
Ever-lived on a farm		0.44	
Yes	1.76 (0.42 – 7.41)		*
No	Ref.		
Lived on a farm in first year of life		0.21	
Yes	1.98 (0.68 – 5.71)		1.55 (0.53 – 4.52)
No	Ref.		
Hereditary:			
Familial History of cancer:			
Father ever had cancer		0.52	
Yes	1.31 (0.58 – 2.98)		*
No	Ref.		
Mother ever had cancer		0.32	
Yes	0.60 (0.22 – 1.63)		*
No	Ref.		
Sibling(s) ever had cancer		0.30	
Yes	1.63 (0.65 – 4.12)		*
No	Ref.		
Occupational Exposures:			
At work, ever exposed to:			
Asbestos dust	0.53 (0.07 – 3.92)	0.54	
Diesel fumes	1.11 (0.51 – 2.42)	0.78	
Fungicides (to treat grain)	1.36 (0.64 – 2.89)	0.42	
Grain dust	0.83 (0.37 – 1.83)	0.64	
Pesticides (to kill plants and insects)	0.95 (0.44 – 2.04)	0.90	
Livestock	1.02 (0.48 – 2.19)	0.95	
Mine dust	1.31 (0.31 – 5.48)	0.72	*
Molds	0.71 (0.31 – 1.62)	0.41	*
Oil/Gas well fumes	1.71 (0.77 – 3.80)	0.19	1.81 (0.77 – 4.29)
Radiation	2.24 (0.85 – 5.89)	0.10	1.69 (0.62 – 4.64)
Stubble smoke	1.35 (0.64 – 2.87)	0.43	*
Solvent fumes	1.4 (0.66 – 2.99)	0.38	*
Welding fumes	0.83 (0.38 – 1.80)	0.63	*
Wood dust	0.99 (0.45 – 2.13)	0.98	*
Covariates			
Age (yrs.)		0.01	
50-59	Ref.		Ref.
60-69	1.19 (0.39 – 3.69)		1.13 (0.37 – 3.44)
70+	3.72 (1.45 – 9.54)		3.22 (1.21 – 8.58)

BMI (kg/m^2)		0.89	
Normal (0-<25)	Ref.		Ref.
Overweight (25-30)	0.91 (0.35 – 2.33)		0.97 (0.36 – 2.64)
Obese (>30)	1.12 (0.42 – 3.01)		1.29 (0.43 – 3.89)
Education		0.54	
≤ Grade 12	Ref.		*
> Grade 12	0.77 (0.34 – 1.77)		
Marital status		0.59	
Married/common law/living together	0.77 (0.29 – 2.02)		*
Widowed/Divorced/separated/single/never married	Ref.		
Sex		0.57	
Male	Ref.		Ref.
Female	0.81 (0.38 – 1.72)		1.01 (0.43 – 2.37)

5.6 Conclusion: Summarizing methods used for data analysis

The methods used to answer the research questions in this thesis are summarized in Table 5.16.

Table 5. 16 Summarizing methods used for data analysis

	Methodology	Methods of parameter estimation	Software	Procedure/Command
Research question 1				
Crude prevalence	X individuals in population with CRC at a given time)/(population)	*	Manually by hand	*
Adjusted prevalence (for cross-sectional and longitudinal data)	Generalized estimating equation (GEE) with an exchangeable correlation matrix	Multivariate quasi-likelihood	SAS SPSS	GENMOD GENLIN
Research question 2				
Crude incidence	Newly diagnosed cases/population at risk	*	Manually by hand	*
Adjusted Incidence	Cox's proportional hazard model	Partial-likelihood	STATA	STSET STCOX VCE(CLUSTER _) STPHTEST

CHAPTER 6 – DISCUSSION

6.1 Introduction

The Saskatchewan Rural Health Study (SRHS) was a prospective cohort study with a baseline survey conducted in 2010 and a four-year follow-up in 2014. Individuals aged 50 years or older were selected because that is the population we considered to be at risk and also the youngest CRC case was aged 50 years. Baseline and follow-up data sets were first analyzed separately to determine; (1) the crude prevalence of CRC and (2) risk factors for the adjusted prevalence of CRC at the two time-points. The data sets were then combined to determine the longitudinal changes in CRC prevalence over the period 2010 to 2014. Crude and adjusted incidences of CRC incidence were also estimated using the SRHS data set. Data analysis contained in this thesis was based on Health Canada's "Population Health Framework" (PHF) which proposes that individual and contextual factors may interact to produce different risk levels of an adverse health outcome [138]. The following sections summarize/discuss the results of the current study. Section 6.1 and 6.2 respectively discuss results for the prevalence and longitudinal changes in CRC prevalence as well as risk factors associated with them. A similar discussion for CRC incidence is provided in Section 6.3. The study strengths and limitations are stated in Section 6.3 while the conclusion is provided in Section 6.4.

6.2 Prevalence

In total 62 CRC cases were reported at baseline and 66 at follow-up. Majority of the CRC cases were farm resident at both time-points. The crude prevalence of CRC cancer in the current study was 1.1% at baseline and 1.7% at follow-up. A meta-analysis by Heitman et al reveals that crude prevalence rates of CRC range from 0 to 1.68% [142]. Our crude prevalence rate of CRC is consistent with this finding. Johnson et al [148] and Mehran et al [151] both reported a crude prevalence rate of 1.1% which matches the estimate at baseline in the current study. Similarly, the crude prevalence of CRC at follow-up (1.7%) reported in this thesis is similar to the 1.68% reported by Disario et al [150]. Our observations show a slightly higher crude prevalence of CRC as compared to recently reported rates several major North-American studies. Spellman et al [143], Kim et al [144], Rex et al [145], Stevens et al [146], Prajapati et al [147], Imperiale et al [149] and Pickhardt et al [152] reported crude CRC prevalence rates as 0.48, 0.13, 0.20, 0.37, 0.39, 0.70, and 0.16 respectively. The reason for the difference in crude prevalence estimates might be due to differences in the definition of CRC and the age distribution of study participants. For example, Kim et al [144] defined CRC based on the detection of advanced neoplasia (i.e. adenomas and carcinomas) as well as the number of identified polyps while Imperiale et al [149] identified CRC cases by first subjecting study participants to a standard Hemoccult II test and then colonoscopy. However, in the current study, CRC status was self-reported doctor diagnosed. On the other hand, Heitman et al [142] reported a crude prevalence of CRC of 0.7% in a study population aged 65 years or older while in the current study, participants aged 50 or older were analyzed. We discuss below risk factors for CRC.

6.2.1 Risk factors for CRC prevalence and longitudinal changes in prevalence

6.2.1.1 Environmental/Contextual factors

At baseline, the prevalence of CRC among farm residents was 0.8% among farm residents and 1.4% among non-farm residents. At follow-up, the prevalence of CRC among farm residents increased to 1.3% and that among non-farm residents increased to 2%. Although numerical differences in the crude prevalence of CRC was observed between farm and non-farm residents, a statistically significant difference could not be established (p>0.05). In the univariate analysis, farm residence was significantly associated with lower CRC risk both at baseline and at follow-up. A similar relationship was found between farm residence and CRC prevalence at baseline in the multivariable analysis after accounting for other covariates even though not significant (OR = 0.64, 95% CI: 0.30, 1.34). We hypothesized that farm residence would increase the risk of CRC cancer. After the multivariable adjustment, our results did not support our original hypothesis.

Quadrant was found to influence CRC risk. CRC was found to be more prevalent in the eastern part of the Province of Saskatchewan as compared to the western part. The reason for this trend is remains unknown although we suspect it is due to the fact that majority (51.92%) of the older adults live in Eastern part as compared to those living in the Western part (48.08%). This reflects differences in the distribution of CRC prevalence within the same geographic region. This finding was supported by other studies that reported greater variations in CRC prevalence within the same geographical regions [153]. For instance, in Europe, an estimated overall CRC prevalence among men aged 50-74 years varied from 0.19% in Albania to about 1.01% in Slovenia [153]. Higher prevalence rates were reported among men in The Czech Republic (1.13%), Slovakia (1.19%), and Hungary (1.27%), and among women, for Denmark (0.66%), Norway, and The Netherlands (0.64%, both) [153]. In the multivariable analysis, there was a higher risk of CRC in the Eastern quadrants than the western quadrants although this relationship was not statistically significant. The reason for these differences is not completely understood. To the best of our knowledge, this is the first study in Canada to explore the prevalence of CRC in rural Saskatchewan.

CRC was more prevalent in non-smoking households than in-home smoking households. Having a householder member who smoked within the household was associated with a lower risk of CRC at baseline and follow-up (OR = 0.66 and OR = 0.73 respectively). This relationship was not significant in the univariable analysis. Our findings contradicted findings in recent studies that linked exposure to second-hand smoke (SHS) to a higher risk of developing CRC [156 – 158]. In all three studies, higher risks were reported among never smokers who were exposed to higher volumes of SHS. It is been argued that carcinogen from cigarette and tobacco smoke can reach the bowel through circulation after transoral uptake [159] or by directly inhaling smoke from cigarette [160].

In the current study, we observed that the consumption of shallow well reduced the risk of CRC although this was not significant in the multivariable analysis (OR = 0.89, 95% CI: 0.30 – 2.52). This relationship contradicted the findings of one study [161]. Zhou et al reported that the consumption of "surface water" was a risk factor for CRC [161]. Our observation also showed that drinking water from rivers, ponds, and/or dugouts was associated with an increased the risk of CRC though this relationship was not statistically significant. This was supported by findings from different studies including Zhou et al [161]. In [161], the authors reported higher CRC risk for people who drink river (RR = 7.94) and pond (RR = 7.70) water. This could be due to the microcystins (a blue-greenish algal toxin) in water from ponds, lakes, and dugouts that causes CRC [161]. The positive detection rate (>59pg/ml) of microcystin in pond and river water is reported to be 17.14% and 36.23% respectively.

6.2.1.2 Lifestyle/individual factors

After the multivariable adjustment, a dose-response relationship was found between CRC risk and smoking. Ex- and current smokers were 1.30 and 1.94 times more likely to be diagnosed with CRC as compared to non-smokers. Our implication of smoke as a risk factor for CRC risk was supported by similar studies in the literature. For instance, Huxley et al. [109] and Tsoi et al. [112] reported a 16% and 20% increased CRC risk among current smokers as compared to non-smokers respectively. Liang et al. reported a 50% increased risk among smokers as compared to non-smokers [113].

6.2.1.3 Early life exposures

In a univariable analysis, we observed that individuals who ever lived on a farm or lived their first-year of life on a farm at follow-up were associated with higher prevalence of CRC. After adjusting for other variables in a multivariable model, individuals who lived their first year of life on farm were still associated with increased CRC prevalence although the relationship was not significant. For the purposes of the current study, to my knowledge, I could not find any previous study that showed the direct relationship between living one's first year of life on a farm and the prevalence of CRC.

6.2.1.4 Occupational Exposures

Our study showed that exposure to grain dust, oil/gas well fumes, and radiation were associated with the prevalence of CRC in a univariable analysis. However, when we adjusted for variables in the model, only exposure to grain dust and radiation were found to be associated with CRC prevalence still, even though these relationships were not statistically significant. Individuals who were exposed to grain dust were less likely to be diagnosed with CRC than individuals who did not. We could not find a similar study in the literature to support or challenge this finding. However, in a subtle contradiction, Peters et al [162] reported that colorectal cancers were associated with jobs at which fumes, and dust are inhaled especially if jobs were held for longer periods. Radiation was associated with elevated risk of CRC prevalence. Our finding was supported by other studies in the literature including that of the American Gastroenterological Association (AGA), which reported that radiation is a risk factor in the natural history of CRC [163].

6.2.1.5 Family history of Cancer

In a univariable analysis, we observed that individuals whose first-degree relatives (FDR) (i.e. father, mother, and sibling) had a previous history of cancer were more likely to be diagnosed with CRC than individuals with FDR who had no such history both at baseline and follow-up. This agreed with the findings in other studies. One study used data from the Swedish Family Cancer database and reported that a parent with a history of cancer (especially CRC) was associated with a 2-fold risk of being diagnosed with CRC, and the risk triples if the parent was diagnosed before 60 years [164]. After adjusting for other variables, only persons with mothers having a history of cancer were significantly associated with CRC prevalence. Similar to the findings of this study, a Chinese study of women reported a significant association between CRC prevalence and having a parent with a history of CRC [165]. However, no association was found between CRC prevalence and having a sibling who has a history of CRC [165].

6.2.1.6 Other risk factors (Increasing age and gender)

After a multivariable adjustment, our observations revealed a significant ($p<0.05$) dose-response relationship between age and CRC prevalence at baseline. This finding agrees with findings reported in other studies [84-86]. In the current study, 100% of CRC cases were 50 years older, which is in tandem with a report that over 90% of CRC cases occur in individuals aged 50 years or older [86]. However, there is an emerging body of knowledge reporting that CRC prevalence is increasing for younger age cohorts (i.e. among individuals aged 40 or younger). The mean age at diagnosis was 69 years at baseline and 71 at follow-up (males and females combined). This is similar to the mean age at diagnosis of CRC reported by Howlader et al [166] to be 68 years in males and 72 years in females. In a univariable analysis, it was observed that females were more likely to be diagnosed with CRC than males. However, in a follow-up survey, this association reversed. The reasons for the gender differences are not fully understood but are reported to reflect variations in exposure to CRC risk factors and the impact of female-specific hormones [92].

6.3 Incidence

In the current study, we observed an incidence rate of 1.98 per 1,000 person-years (i.e. 27/13,632 total time under observation and at risk), resulting in a cumulative incidence of 0.8% during the study period (2010 - 2014). Generally, our risk estimate was lower than a hazard ratio of 1.7 associated with incident CRC in a similar cohort study [171]. In the Health Professional Follow-up and the Nurses' Health studies, a relative risk of incident cancer was reported at 1.7 which is numerically higher than the risk estimate in the current study [172]. Our study shows that CRC incidence is low in a province where farming is the predominant activity, which is in sharp contrast with the finding by Carrozza et al. [131] reported that CRC incidence tends to be high in regions with greater farming activity.

Our study also shows that the incidence of CRC is low in an under-developed region. This was consistent with the findings in a similar study where the crude incidence of CRC in SSA (a developing region) is reported to be about 4.04 per 100,000 persons [77]. Several studies of cancer incidence have consistently reported a lower risk for tobacco- and cigarette-related cancers related [167-169] (such as CRC) among agricultural populations. Reduced CRC incidence has been linked to these favorable risk factor profiles which include high levels of physical activity [167-169]. However, in the current study, physical activity was not significant in the univariable analysis. Given that our study showed a relatively small increase in CRC incidence in rural residents of Saskatchewan aged 50 years or older, more aggressive colonoscopy and other forms of CRC detections would likely increase CRC incidence rates.

6.3.1 Risk factors for CRC Incidence

6.3.1.1 Environmental/contextual factors

Our observation showed that quadrant was associated with CRC incidence in both the univariable and multivariable Cox's proportional hazard model. When adjusted for other variables in the Cox model, it was observed that the adjusted hazards ratio for CRC incidence was highest in NE and least in NW. These associations were however not statistically significant. Household use of natural gas was significantly associated with CRC incidence in univariable Cox regression analysis. However, after adjusting for other variables in the multivariable models, natural gas was no longer a determinant of CRC risk. Studies implicating household use of natural gas as a risk factor for CRC incidence could not be found.

6.3.1.2 Lifestyle/individual factors

Lifestyle factors included in the current study were smoking status and alcohol consumption. It was observed that none of these was significantly associated with CRC incidence in the univariable Cox's regression analysis and therefore excluded from all multivariable analysis. However, the authors of European Prospective Investigation into Cancer and Nutrition (EPIC) study found an increased hazard ratio of alcohol consumption on the incidence of CRC [123]. In 2009, the IARC announced that smoking cigarette causes CRC [111]. In the current study, we found that being a current smoker was associated with higher hazards of CRC incidence, this was not statistically significant. Our implication of CRC to be associated with cigarette smoking was consistent with findings from a similar study [108].

6.3.1.3 Early life exposures

The role of exposure to farm environment (especially at an early age) in CRC incidence was investigated. We found that farm residents had a lower hazard of CRC incidence in the univariable analysis. However, after the multivariable adjustment, we observed that farm residence was no longer a significant predictor of CRC incidence. Our finding that farm residents had a lower hazard of CRC incidence was consistent with findings in a similar study [130]. In a study of cancer incidence among a cohort of female farm residents in New York State, Wang et al. reported that persons who have farm jobs and rural residents were associated with significantly lower CRC risk as compared to those with non-farm jobs and urban residents respectively [130]. However, another study found increasing hazards of CRC among farm residents [170]. This could be explained that farm residence is proxy for exposure to pesticides which causes CRC [130].

6.3.1.4 Occupational exposure

Our observation showed that exposure to oil/gas well fumes and radiation were associated with higher hazards of CRC in the Cox univariable regression models. These occupational exposures still recorded higher hazard ratio for CRC incidence in the multivariable analysis even though statistical significance could not be established. Our findings should be interpreted with a caution since the literature implicating exposure to oil/gas well fumes and radiation as risk factors for CRC incidence is limited. Nevertheless, the possibility of an existential association between certain forms of radiation and oil/gas well fumes deserves further evaluation. Our study did not find exposure to pesticides to be associated with CRC incidence. However, pesticides exposure has been implicated to be positively associated with CRC incidence [128].

6.3.1.5 Family history of Cancer

Little is known regarding the change in risk pathways conferred by a family history of CRC [171] because most studies investigating family history-related CRC incidence are retrospective and case-control studies [173-175]. Our prospective cohort study observed that family history of cancer was not associated with CRC incidence in Cox's proportional hazards regression model. Our findings of no association were inconsistent with the findings of a similar prospective cohort study [171]. In a randomized controlled prospective CRC study, Schoen et al [171] reported significantly higher hazards ratio of CRC incidence among individuals with a family history of cancer. The authors reported a dose-response relationship between an increasing number of FDRs and CRC incidence [172], a relation that was not explored in the current study.

Park et al [176] found a family history of CRC to be a determinant of advanced adenoma multiplicity. They reported that individuals with FDRs with history CRC were 2 times at a higher risk of developing CRC or precursor lesions adenoma polyps as compared to individuals with FDRs with no such history [176]. The risk is even higher (i.e. 4 times) for individuals with more than one FDRs with a history of CRC [176]. This could be explained with the fact that a family history of CRC may convey genetic susceptibilities that facilitates the formation of new lesions and/or enhances the transition from adenomas to carcinomas [176].

6.3.1.6 Other risk factors (Increasing age and BMI; gender)

In the present study, we observed a dose-response relationship between CRC incidence and age and BMI. After a multivariate adjustment in a Cox's proportional hazard regression model, individuals aged 60–69 years were associated with a high hazard of CRC incidence. An even higher hazard ratio was observed among persons aged 70 or older. Our findings regarding a dose-response relationship were inconsistent with findings from two recent studies. Siegel et al (2019) and Davis et al (2011) reported decreasing CRC incidence among individuals aged 50 or older [177-178]. However, there is mounting evidence that CRC incidence was increasing in younger adults. Recent data from the United States Surveillance, Epidemiology, and End Results Reporting (SEER) database show that CRC incidence is increasing among the under-40 age cohort [177-178] which was not supported by the current study because the minimum age at diagnosis was 50. There is also emerging knowledge of increasing CRC incidence among young adults aged 20-39 years even though the absolute increase is lower than among older adults [179]. Although the rising incidence of CRC among the latter group was not the focus of the current study, we concede a possibility of this phenomenon in various parts of Saskatchewan.

Obese (BMI >30) individuals were associated with a higher hazard ratio of CRC incidence (although not statistically significant) after adjusting for other variables in a multivariable Cox's regression model. Implicating higher BMI as a risk factor for CRC incidence in the current study was consistent with the findings of a similar cohort study of CRC incidence among Swedish women [180]. Mechanisms that are potentially responsible for the association between excess body weight and CRC risk are completely not known [121]. However, certain biologic correlates of obesity such as decreased insulin sensitivity and increased circulating estrogen, which are associated with excess body adiposity and not overall body adiposity [94] have been implicated. However, some studies report that overweight- and obese-related CRC risk does not merely result from increased energy intake but may be due to individual differences in metabolic efficiency [94]. Recent studies suggest that persons who burn energy efficiently may be associated with lower CRC risk [181].

After adjusting for other variables and covariates in the Cox's PH multivariable regression model, female gender was associated with a higher hazard ratio as compared to male gender although this was not statistically significant. The reason for this is not completely known but may be due to the intricacies of female hormones [182].

6.4 Study Strengths and limitations

6.4.1 Study Strengths

A major strength of this study is the extensive information collected on individual and contextual factors including important covariates using mail-out self-administered questionnaires at baseline and follow-up surveys [134, 137]. The content of the baseline questionnaire was optimized using a pilot survey [134]. Based on the pilot survey, some questions were modified in the baseline questionnaire before it was mailed to respondents to be self-administered. A mailed questionnaire was the most feasible and prudent option because the sampled population lived in the farthest corners of the Province of Saskatchewan (See Appendix D). Several authors have discussed the practicality of mail questionnaires and concluded that, with the increasing cost of interviews and data collection in surveys, a mail questionnaire is the best option for collecting data in far and wide geographical regions [136, 183]. The study was based on Health Canada's "Population Health Framework" (PHF) [138], which was successful in a similar cohort study [184].

The SRHS employed an interdisciplinary approach from the conceptualization of the study until the completion of the follow-up survey. Geographers, nurses, project managers, epidemiologists, biostatisticians, etc. were brought together as part of the SRHS team. The Rural municipality and small town's council members were involved in every step of the decision-making process. The composition of the team led to rich discussions and feedback based on a one-health paradigm as evidenced by the overall success of the study and response rate (42%) for a mail-out survey. At baseline, the participation among farm and non-farm rural residents were approximately equal, 42.2% and 41.9% respectively.

Our study identified risk factors for the prevalence, incidence as well as the risk factors for the longitudinal changes in the prevalence of CRC in a rural population. Our results show that age had a dose-response relation between longitudinal changes in CRC prevalence and incidence. Majority of our results were consistent with most results from already published literature.

The prevalence of CRC was based on the question, “Has a doctor of PCG ever said you have……colorectal cancer?”, a criterion widely used in medical studies [185]. This self-reported pathological diagnosis would reduce the possibility of falsely classifying persons who had polyps or in situ cancer as having CRC.

6.4.2 Study Limitations

Our current study was not without limitations. One of the main limitations of this study is the small number of CRC cases in both the baseline (62/5,599) and follow-up (63/3,933) surveys. As a result of this sample size, the power of statistical tests and models was low and might not be able to detect real differences where indeed there are true differences. This may be evidenced by the many statistically not-significant associations found in this study and so many interactions could not be investigated.

Although ethnicity or race is a widely published risk factor associated with CRC prevalence and incidence, this relationship could not be explored in this study because 97% or our study population were Caucasians. As a result, the findings of this study are not a true representative of the rural population of Saskatchewan and/or Canada that have more diverse ethnicity. For the purposes of external validity, the results of this study can only be generalized to similar rural settings in Canada and elsewhere where farming is the predominant occupation.

Regarding occupational history and exposures, there is a possibility of recall bias because participants were required to go back in time to remember whether or not they were exposed to radiation, molds, grain dust, fungicides, asbestos dust, wood dust, diesel fumes, etc. This could result in the misclassification and underestimation of CRC prevalence. On the other hand, this study may have overestimated CRC prevalence if individuals without the disease (i.e. non-cases) were less likely to return a completed questionnaire. The response rate (i.e. 42%) was low in this study. In addition, incidence and prevalence rates were calculated based on self-reported CRC cases. We did not link our study data to the Saskatchewan Cancer Registry (SCR) to confirm the diagnosis of CRC. The response rate (i.e. 42%) was also low in this study both at baseline and follow-up and this could affect the prevalence and incidence. It is also important to mention that the prevalence and incidence analysis and conclusions contained in this thesis is only for a cohort of people aged 50 or older and cannot be generalized to people younger than 50.

The BMI variable in this study was derived from self-reported height and weight of respective participants. This results in the possibility of heights being over-reported and weights under-reported. Nonetheless, a self-reported BMI is shown to be a valid and sufficient estimate of disease prevalence among obese and overweight individuals [186]. The presence of mold in a household was also self-reported and required independent and objective confirmation, which was not done in this study.

Due to the low prevalence and incidence of CRC in rural Saskatchewan, in future, we would consider using a case-control design to analyze the data. In addition, we would also consider including people younger than 50 years in all analysis since it has been reported that CRC risk is increasing in this age cohort. A spatial analysis would also be done in future to understand the real reasons for the variations in the prevalence and incidence rates across the quadrants of Saskatchewan.

6.5 Conclusion

The overall crude prevalence of CRC was 1.1% at baseline and 1.7% at follow-up. When stratified by farm and non-farm residence, the crude prevalence of CRC was almost two times lower among farm residents than among non-farm residents (0.8% vs. 1.3%) at baseline. Among farmers who self-reported doctor-diagnosed CRC, the highest prevalence occurred in SE (1.4%) while the least occurred in SW (0.3) and NW (0.3) respectively. Among non-farm residents, CRC prevalence was highest in NW (1.5%) and least in SW (1.1%). Overall, CRC was more prevalent in the eastern part of the quadrant [SE (1.3%) and NE (1.3%)] as compared to the western part [(SW (0.8%) and NW (1.0%))] at baseline.

The crude incidence rate of 1.98 per 1,000 person-years (i.e. 27/13,632 total time under observation and at risk), resulting in a cumulative incidence of 0.8% during the study period (2010 - 2014).

Prevalence and longitudinal changes in prevalence as well as incidence of CRC among farm and non-farm rural residents appear to depend on a complex combination of individual and contextual factors.

REFERENCES

1. Stoltzfus, J. (2011). Logistic Regression: A Brief Primer. Academic Emergency Medicine, 8(10),1099-1104.
2. Pahwa, P. (2000). Statistical modelling of longitudinal lung function data (Doctoral dissertation). University of Saskatchewan, Saskatoon, Canada.
3. Liang, K., & Zeger, S. (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, 73(1), 13-22.
4. Zeger, S., & Liang, K. (1986). Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics, 42(1), 121-130.
5. Wacholder, S. (1986). Binomial regression in GLIM: Estimating risk ratios and risk differences. American Journal of Epidemiology, 123(1), 174-84.
6. Traissac, P., Martin-Prével, Y., Delpeuch, F., & Maire, B. (1999). Logistic regression vs other generalized linear models to estimate prevalence rate ratios. Revue D'epidemiologie Et De Sante Publique, 47(6), 593-604.
7. Martuzzi, & Elliott. (1998). Estimating the Incidence Rate Ratio in Cross-Sectional Studies Using a Simple Alternative to Logistic Regression. Annals of Epidemiology, 8(1), 52-55.
8. Lipsitz, S., & Zhao, L. (1994). Jackknife Estimators of Variance for Parameter Estimates from Estimating Equations with Applications to Clustered Survival Data. Biometrics, 50(3), 842-846.
9. Rubin, D. (1976). Inference and Missing Data. Biometrika, 63(3), 581-592.
10. Laird, N. M., and Ware, J. H. (1982). Random effects models for longitudinal data. Biometrics, 38:963-974
11. George, B., Seals, S., & Aban, I. (2014). Survival analysis and regression models. Journal of Nuclear Cardiology, 21(4), 686-694.
12. Kaplan, E.L. & Meier, P. (1958). Non-parametric estimation from incomplete observations, Journal of the American Statistical Association 53, 457–481, 562–563.
13. Savage IR. Contributions to the theory of rank order statistics: The two-sample case. Ann Math Stat.1956; 27(3):590–615.
14. Kleinbaum, D., Klein, Mitchel. author, & SpringerLink. (2005). Survival Analysis: A Self-Learning Text (Second ed., Statistics for Biology and Health).
15. Cox, D. (1972). Regression Models and Life-Tables. Journal of the Royal Statistical Society. Series B (Methodological), 34(2), 187-220.
16. Wei, L. (1992). The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Statistics in Medicine, 11(14 ‐ 15), 1871-1879.
17. Wei, L., Ying, Z., & Lin, D. (1990). Linear Regression Analysis of Censored Survival Data Based on Rank Tests. Biometrika, 77(4), 845-851.
18. Prentice, R. (1983). ‘Linear rank tests with right censored data’. Biometrika, 70(1), 304.
19. Burton, P., Gurrin, L., & Sly, P. (1998). Extending the simple linear regression model to account for correlated responses: An introduction to generalized estimating equations and multi-level mixed modelling. Statistics in Medicine, 17(11), 1261-1291.
20. Diggle, P.J., Liang, K.Y. and Zeger, S.L. (1994) Analysis of Longitudinal Data. Oxford University Press, New York.
21. Wang, Annie, Wang, Mo-Jin, & Ping, Jie. (2015). Clinicopathological features and survival outcomes of colorectal cancer in young versus elderly: A population-based cohort study of SEER 9 registries data (1988-2011). Medicine, 94(35), E1402.
22. Colorectal Cancer - Overview. American Society of Clinical Oncology (ASCO). (2013, September). *Cancer.Net.* Alexandria, VA.: American Society of Clinical Oncology (ASCO). Available at: http://www.cancer.ca/en/cancer-information/cancer-type/colorectal/colorectal-cancer/?region=on#ixzz5hu7U0cnz
23. Statistics Canada. Population, urban and rural, by province and territory - Canada, census of population 1851-2006. 2009; Available at: http://www.statcan.gc.ca/tablestableaux/sumsom/l01/cst01/demo62a-eng.htm. (Accessed October 8, 2018 at 01:41pm).

24. Statistics Canada. Population, urban and rural, by province and territory- Saskatchewan, census of population 1851-2006. 2009; Available at: http://www.statcan.gc.ca/tablestableaux/sumsom/l01/cst01/demo62i-eng.htm (Accessed October 8, 2018 at 01:41pm).
25. Monson, R. (1990). Occupational Epidemiology. 2nd Ed. Boston: CRC Press.
26. Hosmer, David W., Lemeshow, Stanley, & Sturdivant, Rodney X. (2013). Model-Building Strategies and Methods for Logistic Regression. In Wiley Series in Probability and Statistics (pp. 89-151). Hoboken, NJ, USA: John Wiley & Sons.
27. Breslow, N. (1974). Covariance Analysis of Censored Survival Data. Biometrics, 30(1), 89-99.
28. Lee, J., & Chia, K. (1993). Estimation of prevalence rate ratios for cross sectional data: An example in occupational epidemiology. British Journal of Industrial Medicine, 50(9), 861-862.
29. Singer, Julio & Andrade, Dalton. (2000). Analysis of longitudinal data. Handbook of Statistics. 18. 115-160. 10.1016/S0169-7161(00)18007-1.
30. Box, G. (1950). Problems in the Analysis of Growth and Wear Curves. Biometrics, 6(4), 362-389.
31. Geisser, S., & Greenhouse, S. (1958). An Extension of Box's Results on the Use of the F Distribution in Multivariate Analysis. The Annals of Mathematical Statistics, 29(3), 885-891.
32. Rao, C. (1959). Some Problems Involving Linear Hypotheses in Multivariate Analysis. Biometrika, 46(1/2), 49-58.
33. Rao, C. (1965). The Theory of Least Squares When the Parameters are Stochastic and Its Application to the Analysis of Growth Curves. Biometrika, 52(3/4), 447-458.
34. Potthoff, R., & Roy, S. (1964). A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems. Biometrika, 51(3/4), 313-326.
35. Grizzle, J., & Allen, D. (1969). Analysis of Growth and Dose Response Curves. Biometrics, 25(2), 357-381.
36. Prentice, R., & Zhao, L. (1991). Estimating Equations for Parameters in Means and Covariances of Multivariate Discrete and Continuous Responses. Biometrics, 47(3), 825-839.
37. Liang, K., Zeger, S., & Qaqish, B. (1992). Multivariate Regression Analyses for Categorical Data. Journal of the Royal Statistical Society. Series B (Methodological), 54(1), 3-40.
38. Fitzmaurice, G., Laird, N., & Rotnitzky, A. (1993). Regression Models for Discrete Longitudinal Responses. Statistical Science, 8(3), 284-299.
39. Diggle, P., Heagerty, Patrick, Liang, Kung-Yee, & Zeger, Scott L. (2002). Analysis of longitudinal data (2nd ed., Oxford statistical science series; 25). Oxford; New York: Oxford University Press.
40. Raghu Bahadur, Raj. (1959). A representation of the joint distribution of responses to n dichotomous items. Studies in Item Analysis and Prediction, Stanford Mathematical Studies in the Social Sciences IV. 27.
41. Bishop, Y.M.M., et al. (1975) Discrete Multivariate Analysis: Theory and Practice. Mass MIT Press, Cambridge, 18-37.
42. Wedderburn, R.W.M. (1974). *Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method.* Biometrika, 61, p. 439-447.
43. McCullagh, P., & Nelder, John A. (1983). Generalized linear models (Monographs on statistics and applied probability (Series)). London; New York: Chapman and Hall.
44. Agresti, A. (2013) Categorical Data Analysis. 3rd Edition, John Wiley & Sons Inc., Hoboken.
45. McCullagh, P., Nelder, J. (1989). Generalized Linear Models, Second Edition. Chapman & Hall. ISBN: 9780412317606
46. Stevens, C., Smith, S., Vrinten, C., Waller, J., & Beeken, R. (2018). Lifestyle changes associated with participation in colorectal cancer screening: Prospective data from the English Longitudinal Study of Ageing. Journal of Medical Screening, 969141318803973.
47. Rabeneck, Paszat, Hilsden, Saskin, Leddin, Grunfeld, . . . Stukel. (2008). Bleeding and Perforation After Outpatient Colonoscopy and Their Risk Factors in Usual Clinical Practice. Gastroenterology, 135(6), 1899-1906.e1.
48. Glanz, Grove, Lerman, Gotay, & Le Marchand. (1999). Correlates of intentions to obtain genetic counseling and colorectal cancer gene testing among at-risk relatives from three ethnic groups. Cancer Epidemiology, Biomarkers

& Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology, 8(4 Pt 2), 329-36.

49. Stürmer, Buring, Lee, Kurth, Gaziano, & Glynn. (2006). Colorectal Cancer After Start of Nonsteroidal Anti-Inflammatory Drug Use. The American Journal of Medicine, 119(6), 494-502.
50. Andersen, P. K. and Keiding, N. (2005). Survival Analysis, Overview. In Encyclopedia of Biostatistics (eds P. Armitage and T. Colton). doi:10.1002/0470011815.b2a11072
51. Westergaard, H. (1925). Modern problems in vital statistics, Biometrika 17, 355–364.
52. Cutler, S.J. & Ederer, F. (1958). Maximum utilization of the life table method in analyzing survival, Journal of Chronic Diseases 8, 699–713.
53. B¨ohmer, P.E. (1912). Theorie der unabh¨angigen Wahrscheinlichkeiten, Rapports, M´emoires et Proc´es – verbaux du 7 e Congr`es International d'Actuaires, Amsterdam 2, 327–343.
54. Peterson, A. (1977). Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions. Journal of the American Statistical Association, 72(360), 854-858.
55. Greenwood, M. (1926). The natural duration of cancer. Reports on Public Health and Medical Subjects 33, 1–26. Her Majesty's Stationery Office, London.
56. Nelson, W. (1969). Hazard plotting for incomplete failure data, *Journal of Quality Technology* 1, 27–52.
57. Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* 14, 945 – 965.
58. Mantel, Nathan (1966). "Evaluation of survival data and two new rank order statistics arising in its consideration". *Cancer Chemotherapy Reports*. 50 (3): 163–70. PMID 5910392.
59. Tsiatis, A., & Davidian, M. (2004). Joint Modeling of Longitudinal and Time-to-Event Data: An Overview. Statistica Sinica, 14(3), 809-834.
60. Rasouli, M., Moradi, G., Roshani, D., Nikkhoo, B., Ghaderi, E., & Ghaytasi, B. (2017). Prognostic factors and survival of colorectal cancer in Kurdistan province, Iran: A population-based study. Medicine, 96(6), E5941.
61. Paraf, F., & Jothy, S. (2000). Colorectal cancer before the age of 40. Diseases of the Colon & Rectum, 43(9), 1222-1226.
62. Hassan, M., Suan, M., Soelar, S., Mohammed, N., Ismail, I., & Ahmad, F. (2016). Survival Analysis and Prognostic Factors for Colorectal Cancer Patients in Malaysia. Asian Pacific Journal of Cancer Prevention: APJCP, 17(7), 3575-81.
63. Yuan, Ying, Li, Mo-Dan, Hu, Han-Guang, Dong, Cai-Xia, Chen, Jia-Qi, Li, Xiao-Fen, . . . Shen, Hong. (2013). Prognostic and survival analysis of 837 Chinese colorectal cancer patients. World Journal of Gastroenterology, 19(17), 2650-9.
64. Sharkas, G., Arqoub, K., Khader, Y., Tarawneh, M., Nimri, O., Al-zaghal, M., & Subih, H. (2017). Colorectal Cancer in Jordan: Survival Rate and Its Related Factors. Journal of Oncology, 2017, 6.
65. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 127:2893–2917.
66. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A (2015) Global cancer statistics, 2012. CA Cancer J Clin 65, 87–108.
67. Ferlay J, Soerjomataram I, Ervik M, et al. GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC Cancer Base No. 11. Lyon, France: International Agency for Research on Cancer, 2013.
68. Douaiher, J., Ravipati, A., Grams, B., Chowdhury, S., Alatise, O., & Are, C. (2017). Colorectal cancer—global burden, trends, and geographical variations. *Journal of Surgical Oncology, 115*(5), 619-630.
69. Favoriti, Pasqualino, Carbone, Gabriele, Greco, Marco, Pirozzi, Felice, Pirozzi, Raffaele Emmanuele Maria, & Corcione, Francesco. (2016). Worldwide burden of colorectal cancer: A review. *Updates in Surgery, 68*(1), 7-11.
70. Khuhaprema, T., & Srivatanakul, P. (2008). Colon and Rectum Cancer in Thailand: An Overview. *Japanese Journal of Clinical Oncology, 38*(4), 237-243.

71. Kaw, Punzalan, Crisostomo, Bowyer, & Wherry. (2002). Surgical pathology of colorectal cancer in filipinos: Implications for clinical practice 1 1 No competing interests declared. Journal of the American College of Surgeons, 195(2), 188-195.
72. Ibrahim, E., Zeeneldin, A., El-Khodary, T., Al-Gahmi, A., & Bin Sadiq, B. (2008). Past, present and future of colorectal cancer in the Kingdom of Saudi Arabia. Saudi Journal of Gastroenterology, 14(4), 178-182.
73. Yazdanpanahi, Nasrin, Salehi, Rasoul, & Kamali, Sara. (2018). RAD51 135G>C polymorphism and risk of sporadic colorectal cancer in Iranian population. Journal of Cancer Research and Therapeutics, 14(3), 614-618.
74. Al-Jaberi, Tareq, Ammari, Fuad, Gharieybeh, Kamal, Khammash, Muhammad, Yaghan, Rami, Heis, Hussein, . . . Al-Omari, Najeh. (1997). Colorectal adenocarcinoma in a defined Jordanian population from 1990 to 1995. Diseases of the Colon & Rectum, 40(9), 1089-1094.
75. Sung, Lau, Goh, & Leung. (2005). Increasing incidence of colorectal cancer in Asia: Implications for screening. Lancet Oncology, 6(11), 871-876.
76. Kuriki, K., & Tajima, K. (2006). The increasing incidence of colorectal cancer and the preventive strategy in Japan. Asian Pacific Journal of Cancer Prevention: APJCP, 7(3), 495-501.
77. Graham A., Davies Adeloye L.G., Theodoratou E. and Campbell H. (2012) Estimating the incidence of colorectal cancer in Sub Saharan Africa: A systematic analysis. Journal of global health 2(2).
78. Katsidzira L., Gangaidzo I.T., Mapingure M.P. and Matenga J.A. (2015) Retrospective study of colorectal cancer in Zimbabwe: Colonoscopic and clinical correlates. World journal of gastroenterology: WJG 21(8), 2374.
79. Irabor D., Arowolo A. and Afolabi A. (2010) Colon and rectal cancer in Ibadan, Nigeria: an update.Colorectal Disease 12(7Online), e43-e49.
80. Dakubo J., Naaeder S., Tettey Y. and Gyasi R. (2010b) Colorectal carcinoma: An update of current trends in Accra. [Carcinome colorectal: Une mise à jour des tendances actuelles de Accra].
81. Murphy, G., Devesa, S., Cross, A., Inskip, P., Mcglynn, K., & Cook, M. (2011). Sex disparities in colorectal cancer incidence by anatomic subsite, race and age. International Journal of Cancer, 128(7), 1668-75.
82. Canadian Cancer Society's Advisory Committee on Cancer Statistics [Internet]. Canadian Cancer Statistics 2017. Toronto: Canadian Cancer Society; 2017 (accessed October 7 2018 at 04:35am). Available from: http://www.cancer.ca/Canadian-Cancer-Statistics-2017-EN
83. Saskatchewan Cancer Agency [Internet]. Canadian Cancer Statistics 2017. Toronto: Canadian Cancer Society; 2017 Available from: http://www.saskcancer.ca/Default.aspx?DN=b6d1ec26-b59d-400a-aa40-0d8c2d3f535e (accessed October 13 2018 at 01:25am).
84. Brenner, Kloor, & Pox. (2014). Colorectal cancer. The Lancet, 383(9927), 1490-1502.
85. Macrae FA (2015) colorectal cancer: epidemiology, risk factors, and protective factors. Wolters Kluwer, UpToDate. Available via http://www.uptodate.com/contents/colorectal-cancer-epidemiology-risk-factorsand-protective-factors. (Accessed October 10 2018 at 5:44am).
86. Ries L., Melbert D. and Krapcho M. (2008) et al. SEER Cancer Statistics Review, 1975- 2005 [based on November 2007 SEER data submission]. Bethesda, MD: National Cancer Institute; 2008.
87. O'Connell J.B., Maggard M.A., Liu J.H. and Etzioni D.A. (2003) Rates of colon and rectal cancers are increasing in young adults. The American surgeon 69(10), 866.
88. Fairley, T., Cardinez, C., Martin, J., Alley, L., Friedman, C., Edwards, B., & Jamison, P. (2006). Colorectal cancer in U.S. adults younger than 50 years of age, 1998–2001. Cancer, 107(S5), 1153-1161.
89. Irby K, Anderson WF, Henson DE, Devesa SS (2006) Emerging and widening colorectal carcinoma disparities between Blacks and Whites in the United States (1975–2002). Cancer Epidemiol Biomark Prev 15:792–797.
90. Swan J, Breen N, Coates RJ et al (2003) Progress in cancer screening practices in the United States: results from the 2000 National Health Interview Survey. Cancer 97, 1528–1540.
91. Janout, V., & Kollárová, H. (2001). Epidemiology of colorectal cancer. Biomedical Papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia, 145(1), 5-10.

92. American Cancer Society. Colorectal Cancer Facts & Figures Special Edition 2005. Oklahoma City, OK: American Cancer Society; 2005; (Accessed October 15, 2018 at 3:16pm).
Available from: http://www.cancer. org/docroot/STT/stt_0.asp.

93. Ren, J., Kirkness, C., Kim, M., Asche, C., & Puli, S. (2016). Long-term risk of colorectal cancer by gender after positive colonoscopy: Population-based cohort study. *Current Medical Research and Opinion, 32*(8), 1367-1374.

94. de Jong, A. E., Morreau, H., Nagengast, F. M., Mathus-Vliegen, E. M. H., Kleibeuker, J. H., Griffioen, G., ... Vasen, H. F. A. (2005). Prevalence of adenomas among young individuals at average risk for colorectal cancer. *American journal of gastroenterology*, *100*(1), 139-143. DOI: 10.1111/j.1572-0241.2005.41000.x

95. Grande, M., Milito, G., Attinà, G., Cadeddu, F., Muzi, M., Nigro, C., . . . Farinon, A. (2008). Evaluation of clinical, laboratory and morphologic prognostic factors in colon cancer. World Journal of Surgical Oncology, 6(1), 98.

96. Atkin, W., Morson, B., & Cuzick, J. (1992). Long-Term Risk of Colorectal Cancer After Excision of Rectosigmoid Adenomas. The New England Journal of Medicine, 326(10), 658-662.

97. Patel, S., & Ahnen, G. (2012). Familial Colon Cancer Syndromes: An Update of a Rapidly Evolving Field. Current Gastroenterology Reports, 14(5), 428-438.

98. Louise E Johns, & Richard S Houlston. (2001). A systematic review and meta-analysis of familial colorectal cancer risk. American Journal of Gastroenterology, 96(10), 2992-3003.

99. Boardman, L., Morlan, B., Rabe, K., Petersen, G., Lindor, N., Nigon, S., . . . Gallinger, S. (2007). Colorectal cancer risks in relatives of young-onset cases: Is risk the same across all first-degree relatives? Clinical Gastroenterology and Hepatology: The Official Clinical Practice Journal of the American Gastroenterological Association, 5(10), 1195-8.

100. Samadder NJ, Curtin K, Tuohy TM, et al. Increased risk of colorectal neoplasia among family members of patients with colorectal cancer: a population-based study in Utah. *Gastroenterology*. 2014, 147:814-821 e815; quiz e815-816.

101. Lutgens, M., Van Oijen, M., Van Der Heijden, G., Vleggaar, F., Siersema, P., & Oldenburg, B. (2013). Declining risk of colorectal cancer in inflammatory bowel disease: An updated meta-analysis of population-based cohort studies. Inflammatory Bowel Diseases, 19(4), 789-99.

102. Munkholm, P. (2003). Review article: The incidence and prevalence of colorectal cancer in inflammatory bowel disease. Alimentary Pharmacology & Therapeutics, 18 Suppl 2, 1-5.

103. Eaden, J., Abrams, K., & Mayberry, J. (2001). The risk of colorectal cancer in ulcerative colitis: A meta-analysis. Gut, 48(4), 526-35.

104. Willett, W. C. (2005). Diet and cancer: An evolving picture. Journal of the American Medical Association, 293(2), 233-234.

105. World Cancer Research Fund and American Institute for Cancer Research. Food, Nutrition, Physical Activity, and the Prevention of Cancer: A Global Perspective. Washington, DC: American Institute for Cancer Research; 2007.

106. Negri E., Franceschi S., Parpinel M. and La Vecchia C. (1998) Fiber intake and risk of colorectal cancer. Cancer Epidemiology Biomarkers & Prevention 7(8), 667-671

107. Song, Garrett, & Chan. (2015). Nutrients, Foods, and Colorectal Cancer Prevention. Gastroenterology, 148(6), 1244-1260.e16.

108. Johnson, C., Wei, M., Ensor, C., Smolenski, J., Amos, E., Levin, D., & Berry, J. (2013). Meta-analyses of colorectal cancer risk factors. Cancer Causes & Control, 24(6), 1207-1222.

109. Huxley, R., Ansary-Moghaddam, A., Clifton, P., Czernichow, S., Parr, C., & Woodward, M. (2009). The impact of dietary and lifestyle risk factors on risk of colorectal cancer: A quantitative overview of the epidemiological evidence. International Journal of Cancer, 125(1), 171-180.

110. Figueiredo, J., Hsu, L., Hutter, C., Lin, Y., Campbell, P., Baron, J., . . . Amos, C. (2014). Genome-Wide Diet-Gene Interaction Analyses for Risk of Colorectal Cancer (Gene-Diet Interactions and Colorectal Cancer Risk). 10(4), E1004228.

111. Secretan, Straif, Baan, Grosse, El Ghissassi, Bouvard, . . . Cogliano. (2009). A review of human carcinogens—Part E: Tobacco, areca nut, alcohol, coal smoke, and salted fish. Lancet Oncology, 10(11), 1033-1034.

112. Tsoi, Pau, Wu, Chan, Griffiths, & Sung. (2009). T1985 Cigarette Smoking and the Risk of Colorectal Cancer: A Meta-Analysis of Prospective Cohort Studies. Gastroenterology, 136(5), A-614-A-615.

113. Liang PS, Chen TY, Giovannucci E. Cigarette smoking and colorectal cancer incidence and mortality: systematic review and meta-analysis. *Int J Cancer* 2009, 124: 2406–15.

114. Zisman A.L., Nickolov A., Brand R.E., Gorchow A. and Roy H.K. (2006) Associations between the age at diagnosis and location of colorectal cancer and the use of alcohol and tobacco: implications for screening. Archives of internal medicine 166(6), 629-634.

115. Polymnia Galiatsatos, & William D Foulkes. (2006). Familial Adenomatous Polyposis. The American Journal of Gastroenterology, 101(2), 385-98.

116. Boyle, T., Keegel, T., Bull, F., Heyworth, J., & Fritschi, L. (2012). Physical Activity and Risks of Proximal and Distal Colon Cancers: A Systematic Review and Meta-analysis. Journal of The National Cancer Institute, 104(20), 1548-1561.

117. Robsahm, T., Aagnes, B., Hjartåker, A., Langseth, H., Bray, F., & Larsen, I. (2013). Body mass index, physical activity, and colorectal cancer by anatomical subsites: A systematic review and meta-analysis of cohort studies. European Journal of Cancer Prevention: The Official Journal of the European Cancer Prevention Organisation (ECP), 22(6), 492-505.

118. Wolin, K., Lee, I., Colditz, G., Glynn, R., Fuchs, C., & Giovannucci, E. (2007). Leisure-time physical activity patterns and risk of colon cancer in women. International Journal of Cancer, 121(12), 2776-81.

119. Moghaddam A.A., Woodward M. and Huxley R. (2007) Obesity and risk of colorectal cancer: a meta-analysis of 31 studies with 70,000 events. Cancer Epidemiology Biomarkers & Prevention 16(12), 2533-2547.

120. Bianchini F, Kaaks R, Vainio H (2002) Overweight, obesity, and cancer risk. Lancet Oncol 3, 565– 574.

121. Ma Y, Yang Y, Wang F, et al. Obesity and risk of colorectal cancer: a systematic review of prospective studies. *PLoS One* 2013; 8: e53916.

122. Nunez, Carlos, Nair-Shalliker, Visalini, Egger, Sam, Sitas, Freddy, & Bauman, Adrian. (2018). Physical activity, obesity and sedentary behaviour and the risks of colon and rectal cancers in the 45 and up study. BMC Public Health, 18(1), 325.

123. Cleves M, Gould WW, Gutierrez RG, Marchenko YU. (2008). An Introduction to Survival Analysis Using Stata. Second. Texas: Stata Press. pp. 7–19.

124. V Bagnardi, M Rota, E Botteri, I Tramacere, F Islami, V Fedirko, . . . C La Vecchia. (2014). Alcohol consumption and site-specific cancer risk: A comprehensive dose–response meta-analysis. British Journal of Cancer, 112(3), 580-593.

125. Fedirko V, Tramacere I, Bagnardi V, et al. Alcohol drinking and colorectal cancer risk: an overall and dose-response meta-analysis of published studies. *Ann Oncol* 2011, 22: 1958–72.

126. Moskal, A., Norat, T., Ferrari, P., & Riboli, E. (2007). Alcohol intake and colorectal cancer risk: A dose–response meta-analysis of published cohort studies. International Journal of Cancer, 120(3), 664-671.

127. American Cancer Society. Colorectal Cancer Facts & Figures 2017-2019. Atlanta: American Cancer Society; 2017. Available online at https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2017-2019.pdf (Retrieved on October 18, 2018 at 10:59pm).

128. Fredriksson, M., Bengtsson, N., Hardell, L., & Axelson, O. (1989). Colon cancer, physical activity, and occupational exposures. A case‐control study. Cancer, 63(9), 1838-1842.

129. Momenyan, S., Ghalane, S., Sarvi, F., Azizi, R., & Kabiri, F. (2017). The Association between Lifestyle, Occupational, and Reproductive Factors and Colorectal Cancer Risk. Asian Pacific Journal of Cancer Prevention: APJCP, 18(8), 2157-2162.

130. Wang Y, Lewis-Michl EL, Hwang S-A, et al (2002). Cancer incidence among a cohort of female farm residents in New York State. *Arch Environ Occup Health*, 57, 561-7.

131. Carozza SE, Li B, Elgethun K, et al (2008). Risk of childhood cancers associated with residence in agriculturally intense areas in the United States. Environ Health Perspect, 116, 559–65.

132. Lo, A., Soliman, A., Khaled, H., Aboelyazid, A., & Greenson, J. (2010). Lifestyle, occupational, and reproductive factors and risk of colorectal cancer. Diseases of the Colon and Rectum, 53(5), 830-7.

133. El-Zaemey, S., Anand, T., Heyworth, J., Boyle, T., Van Tongeren, M., & Fritschi, L. (2018). Case–control study to assess the association between colorectal cancer and selected occupational agents using INTEROCC job exposure matrix. Occupational and Environmental Medicine, 75(4), 290-295.

134. Pahwa, P., Karunanayake, C., Hagel, L., Janzen, B., Pickett, W., Rennie, D., . . . Dosman, J. (2012). The Saskatchewan rural health study: An application of a population health framework to understand respiratory health outcomes. BMC Research Notes, 5(1), 400.

135. du Plessis V, Beshiri R, Bollman RD, Clemenson H: Definitions of “Rural”. Agriculture and Rural Working Paper Series, Working Paper No. 61. Catalogue no. 21-601-MIE- No. 061. Ottawa, Canada: Agricultural Division, Statistics Canada; 2004.

136. Dillman, DA. (1978). Mail and telephone surveys: The total design method. New York: Wiley.

137. Pahwa, Punam, Rana, Masud, Pickett, William, Karunanayake, Chandima P., Amin, Khalid, Rennie, Donna, . . . Dosman, James. (2017). Cohort profile: The Saskatchewan Rural Health Study--adult component. BMC Research Notes, 10(1), 1-7.

138. Canada. Health Canada issuing body. (1994). Strategies for population health: Investing in the health of Canadians.

139. Dosman J, Pahwa P, et al. Saskatchewan Rural Health Study 2009. Canadian Institutes of Health Research Funded Operating Grant MOP-187209-POP-CCAA-11829.

140. Statistics Canada National Population Health Survey Household Component: Documentation for the Derived Variables and the Constant Longitudinal Variables. Available online [http://www.statcan.gc.ca/imdbbmdi/document/3225_D10_T9_V3-eng.pdf]. Accessed on Oct 26, 2018.

141. Rothman, K.J. and S. Greenland, *Modern Epidemiology*. 1998: Lippincott-Raven.

142. Heitman, Ronksley, Hilsden, Manns, Rostom, and Hemmelgarn. "Prevalence of Adenomas and Colorectal Cancer in Average Risk Individuals: A Systematic Review and Meta-analysis." Clinical Gastroenterology and Hepatology 7.12 (2009), 1272-278. Web.

143. Spellman SJ, Bader M, Zogg DI. Outcomes and complications in average risk colon cancer screning in a community hospital. Am J Gastroenterol 2007, 102:1194.

144. Kim DH, Pickhardt PJ, Taylor AJ, et al. CT colonography versus colonoscopy for the detection of advanced neoplasia. N Engl J Med 2007, 357:1403–1412.

145. Rex DK, Lehman GA, Ulbright TM, et al. Colonic neoplasia in asymptomatic persons with negative fecal occult blood tests: influence of age, gender, and family history Am J Gastroenterol 1993, 88:825–831.

146. Stevens T, Burke CA. Colonoscopy screening in the elderly: when to stop? Am J Gastroenterol 2003; 98:1881–1885.

147. Prajapati DN, Saeian K, Binion DG, et al. Volume and yield of screening colonoscopy at a tertiary medical center after change in medicare reimbursement. Am J Gastroenterol 2003, 98:194–199.

148. Johnson DA, Gurney MS, Volpe RJ, et al. A prospective study of the prevalence of colonic neoplasms in asymptomatic patients with an age-related risk. Am J Gastroenterol 1990, 85:969–974.

149. Imperiale TF, Ransohoff DF, Itzkowitz SH, et al. Fecal DNA versus fecal occult blood for colorectal-cancer screening in an average risk population. N Engl J Med 2004, 351:2704–2714.

150. DiSario JA, Foutch PG, Mai HD, et al. Prevalence and malignant potential of colorectal polyps in asymptomatic, average-risk men. Am J Gastroenterol 1991, 86:941–945.

151. Mehran A, Jaffe P, Efron J, et al. Screening colonoscopy in the asymptomatic 50- to 59-year-old population. Surg Endosc 2003, 17:1974–1977.

152. Pickhardt PJ, Choi JR, Hwang I, et al. Nonadenomatous polyps at CT colonography: prevalence, size distribution, and detection rates. Radiology 2004, 232:784–790.

153. Krilaviciute, Agne, Christian Stock, and Hermann Brenner. "International Variation in the Prevalence of Preclinical Colorectal Cancer: Implications for Predictive Values of Noninvasive Screening Tests and Potential Target Populations for Screening." International Journal of Cancer 141.8 (2017), 1566-575. Web.

154. Wang, M. (1991). Nonparametric Estimation from Cross-Sectional Survival Data. Journal of the American Statistical Association, 86(413), 130-143.

155. *A guide to HIV/AIDS epidemiological and surveillance terms*. 2002, Public Health Agency of Canada.

156. Lilla C, Verla-Tebit E, Risch A, Jager B, Hoffmeister M, Brenner H, Chang-Claude J. Effect of NAT1 and NAT2 genetic polymorphisms on colorectal cancer risk associated with exposure to tobacco smoke and meat consumption. Cancer Epidemiol Biomarkers Prev. 2006, 15:99–107.

157. Nishino Y, Tsubono Y, Tsuji I, Komatsu S, Kanemura S, Nakatsuka H, Fukao A, Satoh H, Hisamichi S. Passive smoking at home and cancer risk: a population-based prospective study in Japanese nonsmoking women. Cancer Causes Control. 2001, 12:797–802.

158. Slattery ML, Edwards S, Curtin K, Schaffer D, Neuhausen S. Associations between smoking, passive smoking, GSTM-1, NAT2, and rectal cancer. Cancer Epidemiol Biomarkers Prev. 2003, 12:882–889.

159. Yamasaki E, Ames BN. Concentration of mutagens from urine by absorption with the nonpolar resin XAD-2: cigarette smokers have mutagenic urine. Proc Natl Acad Sci U S A. 1977, 74:3555–3559.

160. Kune GA, Kune S, Vitetta L, Watson LF. Smoking and colorectal cancer risk: data from the Melbourne Colorectal Cancer Study and brief review of literature. Int J Cancer. 1992, 50:369–372.

161. Zhou, L & Yu, D & Yu, H & Chen, K & Shen, G & Shen, Y & Ruan, Y & Ding, X. (2000). Drinking water types, microcystins and colorectal cancer. Zhonghua yu fang yi xue za zhi [Chinese journal of preventive medicine]. 34, 224-6.

162. Peters, R K, D H Garabrant, M C Yu, and T M Mack. "A Case-control Study of Occupational and Dietary Factors in Colorectal Cancer in Young Men by Subsite." *Cancer Research* 49.19 (1989), 5459-468. Web.

163. American Gastroenterological Association Editorials (1983). Available online at https://www.gastrojournal.org/article/S0016-5085(83)80186-3/pdf (Retrieved on April 3, 2019 at 10;27pm)

164. Leu M, Reilly M, Czene K. Evaluation of bias in familial risk estimates: a study of common cancers using Swedish population-based registers. Journal of the National Cancer Institute. 2008 Sep 17;100(18), 1318–1325.

165. Murphy G, Shu X-O, Gao Y-T, et al. Family cancer history affecting risk of colorectal cancer in a prospective cohort of Chinese women. Cancer Causes & Control. 2009 Oct;20(8), 1517–1521.

166. Howlader N, Noone AM, Krapcho M, et al. *SEER Cancer Statistics Review, 1975-2013*. Bethesda, MD: National Cancer Institute, 2016.

167. Acquavella J, Olsen G, Cole P, Ireland B, Kaneene J, Schuman S, Holden L. Cancer among farmers: a meta-analysis. Ann Epidemiol. 1998, 8(1):64–74. doi: 10.1016/S1047-2797(97)00120-8.

168. Laakkonen A, Pukkala E. Cancer incidence among Finnish farmers, 1995-2005. Scand J Work Environ Health. 2008, 34(1):73–79. doi: 10.5271/sjweh.1167.

169. Pukkala E, Martinsen JI, Lynge E, Gunnarsdottir HK, Sparen P, Tryggvadottir L, Weiderpass E, Kjaerheim K. Occupation and cancer - follow-up of 15 million people in five Nordic countries. Acta Oncol. 2009, 48(5):646–790. doi: 10.1080/02841860902913546.

170. *Feychting M, Plato N, Nise G, et al. Paternal occupational exposures and childhood cancer. Environ Health Perspect2001, 109:193–6.*

171. Schoen, Razzak, Yu, Berndt, Firl, Riley, and Pinsky. "Incidence and Mortality of Colorectal Cancer in Individuals with a Family History of Colorectal Cancer." Gastroenterology 149.6 (2015), 1438-445.e1. Web.

172. Fuchs CS, Giovannucci EL, Colditz GA, et al. A prospective study of FH and the risk of colorectal cancer. N Engl J Med 1994, 331:1669–1674.

173. Johns LE, Houlston RS. A systematic review and meta-analysis of familial colorectal cancer risk. Am J Gastroenterol 2001, 96:2992–3003.

174. Hemminki K, Li X. Familial colorectal adenocarcinoma from the Swedish Family-Cancer Database. Int J Cancer 2001, 94:743–748.

175. Slattery ML, Kerber RA. Family history of cancer and colon cancer risk: the Utah population database [published erratum appears in J Natl Cancer Inst 1994 Dec 7;86(23):1802]. J Natl Cancer Inst 1994, 86:1618–1626.

176. Wark, Petra A., Kana Wu, Pieter Van 't Veer, Charles F. Fuchs, and Edward L. Giovannucci. "Family History of Colorectal Cancer: A Determinant of Advanced Adenoma Stage or Adenoma Multiplicity?" International Journal of Cancer 125.2 (2009), 413-20. Web.

177. Siegel, Rebecca L, Kimberly D Miller, and Ahmedin Jemal. "Cancer Statistics, 2019." CA: A Cancer Journal for Clinicians 69.1 (2019), 7-34. Web.

178. Davis, Marcet, Frattini, Prather, Mateka, and Nfonsam. "Is It Time to Lower the Recommended Screening Age for Colorectal Cancer?" Journal of the American College of Surgeons 213.3 (2011), 352-61. Web.

179. Singh, Kathryn, Thomas Taylor, Chuan-Ju Pan, Michael Stamos, and Jason Zell. "Colorectal Cancer Incidence Among Young Adults in California." Journal of Adolescent and Young Adult Oncology 3.4 (2014), 176-84. Web.

180. P Terry, E Giovannucci, L Bergkvist, L Holmberg, and A Wolk. "Body Weight and Colorectal Cancer Risk in a Cohort of Swedish Women: Relation Varies by Age and Cancer Site." British Journal of Cancer 85.3 (2001), 346-9. Web.

181. Boyle P, Langman JS. ABC of colorectal cancer: Epidemiology. BMJ 2000;321(7264), 805–808

182. Issa J.-P.J., Ottaviano Y.L., Celano P., Hamilton S.R., Davidson N.E. and Baylin S.B. (1994) Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. Nature genetics 7(4), 536-540.

183. Hox, J.J.; De Leeuw, E.D. A comparison of nonresponse in mail, telephone, and face-to-face surveys. Qual. Quant. 1994, 28, 329–344. Available online: Available online: https://link.springer.com/article/10.1007/BF01097014 (accessed on April 5 2019 at 19:29GMT). [CrossRef].

184. Pickett W, Day L, Hagel L, Brison RJ, Marlenga B, Pahwa P, et al. The Saskatchewan farm injury cohort: rationale and methodology. Public Health Rep 2008, 123(5):567.

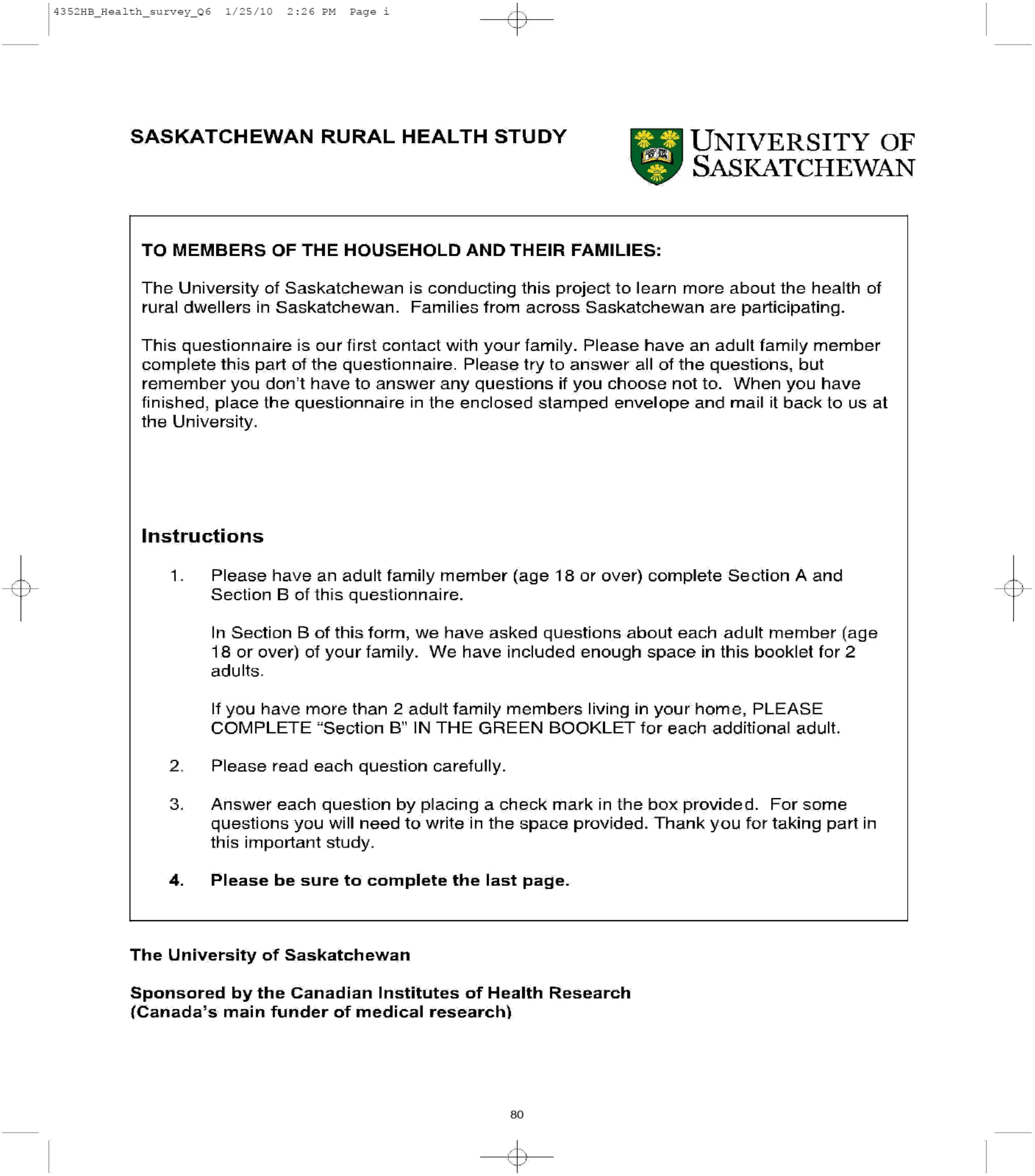
185. Meren M, Jannus-Pruljan L, Loit HM, et al. Asthma, chronic bronchitis and respiratory symptoms among adults in Estonia according to a postal questionnaire. Respiratory Medicine 2001, 95(12):954-964.

186. Fonseca H, Silva AM, Matos MG, et al. Validity of BMI based on self-reported weight and height in adolescents. Acta Paediatrica 2010, 99:83-88.

187. Hogan, Joseph W. "Longitudinal Data Analysis Edited by G. Fitzmaurice, M. Davidian, G. Verbeke, and G. Molenberghs." Biometrics 66.3 (2010), 995-96. Web.

188. Ghosh, Sunita., and University of Saskatchewan, College of Graduate Studies Research. Statistical Modeling of Longitudinal Survey Data with Binary Outcomes (2007). Web.

189. Stata.com. Available online via https://www.stata.com/manuals13/ststcox.pdf (Accessed April 5 2019 at 19:29GMT



4352HB_Health_survey_Q6 1/25/10 2:26 PM Page i

SASKATCHEWAN RURAL HEALTH STUDY

UNIVERSITY OF SASKATCHEWAN

TO MEMBERS OF THE HOUSEHOLD AND THEIR FAMILIES:

The University of Saskatchewan is conducting this project to learn more about the health of rural dwellers in Saskatchewan. Families from across Saskatchewan are participating.

This questionnaire is our first contact with your family. Please have an adult family member complete this part of the questionnaire. Please try to answer all of the questions, but remember you don't have to answer any questions if you choose not to. When you have finished, place the questionnaire in the enclosed stamped envelope and mail it back to us at the University.

Instructions

1. Please have an adult family member (age 18 or over) complete Section A and Section B of this questionnaire.

   In Section B of this form, we have asked questions about each adult member (age 18 or over) of your family. We have included enough space in this booklet for 2 adults.

   If you have more than 2 adult family members living in your home, PLEASE COMPLETE "Section B" IN THE GREEN BOOKLET for each additional adult.

2. Please read each question carefully.

3. Answer each question by placing a check mark in the box provided. For some questions you will need to write in the space provided. Thank you for taking part in this important study.

4. **Please be sure to complete the last page.**

The University of Saskatchewan

Sponsored by the Canadian Institutes of Health Research
(Canada's main funder of medical research)

80

SECTION A YOUR HOME

PLEASE ANSWER THE FOLLOWING QUESTIONS ABOUT YOUR PRIMARY FAMILY HOME - THAT IS THE HOME WHERE YOU LIVE MOST OF THE TIME.

Today's Date: _____/______/_____
(Day / Month / Year)

DEMOGRAPHICS

A-1 Where is your home located?
c Farm
c In town
c Acreage, please specify number of acres ________

A-2 How many people live in your home?
______ Number

A-3 Please list all persons who usually live here including yourself.

Age	Sex		Family Member	
	M c	F c	Yes c	No c
	M c	F c	Yes c	No c
	M c	F c	Yes c	No c
	M c	F c	Yes c	No c
	M c	F c	Yes c	No c
	M c	F c	Yes c	No c

(*IF MORE SPACES ARE REQUIRED CONTINUE ON THE BACK OF THE QUESTIONNAIRE.*)

A-4 How many bedrooms do you have in your home?
______ Number

A-5 Do you own your home?
c Yes
c No
c Don't know

LIVING ENVIRONMENT

A-6 What year was your residence/apartment built (approximately)?

Year______________ Don't know c

A-7 What are the types of fuel sources used to heat your home? **Please check all that apply.**

	Primary	Secondary
c Natural Gas	c	c
c Propane	c	c
c Electricity	c	c
c Fuel oil	c	c
c Coal	c	c
c Geo-thermal	c	c
c Solar energy	c	c
c Wood ➡ If yes, do you use: c Fireplace c Free standing wood stove c Fireplace insert c Outdoor wood stove	c	c
c Other Please specify ______________________________	c	c
c Don't Know		

A-8 Does your heating system have a filter?
c Yes
c No
c Don't Know

A-9 Does your home have air conditioning?
c Yes → **If yes, please check one:**
c Central c Room c Both
c No
c Don't Know

A-10 Is a humidifier or vaporizer used in your home?
c Yes
c No
c Don't Know

A-11 Do you use a dehumidifier in your home?
c Yes
c No
c Don't Know

A-12 On average, how often per month:
do you vacuum carpet? ____ times per month
do you mop smooth floors? ____ times per month
do you dry dust clean? ______ times per month do
you wet dust clean? ______ times per month

A-13 During the past 12 months, has there been water or dampness in your home from broken pipes, leaks, heavy rain, or floods?
c Yes
c No
c Don't Know

Your Home

A-26 In the <u>past 12 months</u>, have you or a family member in your household required immediate 24 hour health care services for a medical emergency?c
Yes

c No → **If No, go to question A-30.**

c Don't know

A-27 In the <u>past 12 months</u>, did you ever experience any difficulties getting immediate 24 hour health care services for a medical emergency for yourselfc or a family member in your household?
Yes

c No

c Don't know

A-28 How far do you travel to receive routine and ongoing medical care?_______ Km

A-29 How far do you travel to receive 24 hour emergency health care services?_______ Km

A-30 How far do you travel to receive medical or surgical specialist services?_______ Km

A-31 On average, how long does it take for an ambulance to arrive at your home in an emergency? ________minutes c Don't Know

OUTDOOR ENVIRONMENT

A-32 Do you have an indoor (barn) intensive livestock operation (building) located near your home?

c Yes → **If Yes, how far?**

c Within 1/4 mile c Greater than 1/4 mile c
No

c Don't know

A-33 Do you have an outdoor feedlot or corrals located near your home?

c Yes → **If Yes, how far?**

c Within 1/4 mile c Greater than 1/4 mile c
No

c Don't know

A-34 Do you have a balestack or bales located near your home?

c Yes → **If Yes, how far?**

c Within 1/4 mile c Greater than 1/4 mile c
No

c Don't know

A-35 Do you have grain bins located near your home?

c Yes → **If Yes, how far?**

c Within 1/4 mile c Greater than 1/4 mile

c No

c Don't know

A-36 Do you have a sewage pond or manure lagoon located near your home?

c Yes → **If Yes, how far?**

c Within 1/4 mile c Greater than 1/4 mile

c No

c Don't know

A-37 What is the <u>**main**</u> source of the water supply for drinking purposes in your home?

c Bottled water
c Deep well water (more than 100 ft)
c Shallow well water (less than 100 ft)
c Spring, river or creek
c Dugout, reservoir
c Lake
c Other source:
Please specify______________________

PLEASE COMPLETE THIS SECTION
IF YOU LIVE ON A FARM.

FARM DEMOGRAPHICS

A-38 From the list below, please check each commodity that is produced for sale on your farm or ranch **(Please check all that apply).**

c Grain crops

c Cattle (beef)

c Cattle (dairy)

c Pigs

c Poultry

c Vegetable/Fruit

c Other:
Please specify______________________

A-39 What is the area of land in your operation that you farmed or ranched last growing season? **(Please exclude land rented to others).**

Grain crops	___ acres
Forage crops	___ acres
Pasture	___ acres
Summerfallow	___ acres
Other	___ acres

A-40 How many of these types of livestock are typically raised on your farm?

No livestock	c
Cattle (beef)	____ number
Cattle (dairy)	____ number
Pigs	____ number
Poultry	____ number
Other	____ number

THIS CONCLUDES SECTION A. PLEASE PROCEED TO SECTION B, ADULT 1(GREEN TAB).

SECTION B INDIVIDUAL QUESTIONS

WE WOULD LIKE TO KNOW ABOUT EACH ADULT FAMILY MEMBER (18 YEARS OR OVER) LIVING IN YOUR HOUSEHOLD. IN THIS BOOKLET, WE HAVE INCLUDED SPACE FOR 2 ADULTS.

> IF YOU HAVE MORE THAN 2 ADULT FAMILY MEMBERS LIVING IN YOUR HOME, PLEASE COMPLETE "Section B" IN THE GREEN BOOKLET FOR EACH ADDITIONAL ADULT.

ADULT 1

NOW, PLEASE ANSWER THE FOLLOWING QUESTIONS ABOUT ADULT # 1.

B-1 Age as of January 1st, 2010: ___________

B-2 Date of birth: MM______ DD____ YY_______

B-3 Sex: Male c Female c

B-4 Highest level of education:

c Less than high school
c Completed high school
c Completed university
c Completed post-secondary education other than above

B-5 What is your ethnic background?

c Caucasian
c First Nation
c Metis
c Other → **Please specify:** ____________

B-6 What is your height? _____cm. **OR** ______ft and in.

B-7 What is your weight? ________Kg. **OR** ______lbs

B-8 What is your marital status? **(Please check only one)**

c Married
c Common law/living together
c Widowed
c Divorced/separated
c Single, never married

RESPIRATORY HEALTH

COUGH

B-9 Do you usually have a cough?

c Yes
c No → **If no, go to question B-12.**

B-10 Do you usually cough like this on most days for 3 consecutive months or more during the year?

c Yes
c No

B-11 For how many years have you had this cough?
_____ years

PHLEGM

B-12 Do you usually bring up phlegm from your chest? c Yes
c No → **If no, go to question B-15.**

B-13 Do you bring up phlegm like this on most days for 3 consecutive months or more during the year?

c Yes
c No

B-14 For how many years have you had trouble with phlegm?
_____ years

WHEEZE

B-15 Does your chest ever sound wheezy or whistling:

	Yes	No
1. When you have a cold?	c	c
2. Apart from colds?	c	c
3. Most days or nights?	c	c

If YES to 1, 2, OR 3, for how many years has this been present? _______________number of years

B-16 Have you ever had an attack of wheezing that has made you feel short of breath?

c Yes
c No

If YES, have you ever required medicine or treatment for the(se) attack(s)?

c Yes
c No

BREATHLESSNESS

B-17 Are you troubled by shortness of breath when hurrying on the level or walking up a slight hill?

c Yes
c No

B-18 Do you have to walk slower than people of your age because of breathlessness?

c Yes
c No

B-19 Do you ever have to stop for breath when walking at your own pace on the level?

c Yes
c No

B-20 Do you ever have to stop for breath after walking about 100 yards (or after a few minutes) on the level?

c Yes

c No

B-21 Are you too breathless to leave the house or breathless on dressing or undressing?

c Yes

c No

ASTHMA

B-22 Have you ever had asthma? c Yes

c No → **If no, go to question B-26.**

B-23 **If Yes to B-22:**

Do you still have it? c Yes c No
Was it confirmed by a doctor? c Yes c No

At what age did it start? ____ age in years

If you no longer have it, at what age did it stop?
____ age in years

B-24 **If yes to B-22**, how many times have you required services for asthma from the following places during the past 12 months?

Hospital inpatient: _________ times
Emergency room outpatient: _________ times
Doctor's office: _________ times

B-25 **If yes to B-22,** which of the following statements best describes your asthma medication use in the past 12 months:

c Never in the past 12 months
c At least once in the past 12 months
c At least once per month
c At least once per week
c Every day

ALLERGIES

B-26 Have you ever had an allergic reaction to any of the following: **(Please check all that apply).**

1. House dust	c Yes	c No
2. Cats	c Yes	c No
3. Dogs	c Yes	c No
4. Grasses	c Yes	c No
5. Pollens	c Yes	c No
6. Molds	c Yes	c No
7. Others,	c Yes	c No

Please specify:____________________

PHYSICAL ACTIVITY

B-27 Do you exercise?

c Yes → **If yes, how many times a week?**
_____times a week

c No → **If no, go to question B-29.**

B-28 How long do you usually exercise?

c Less than 15 minutes
c 15 to 30 minutes
c 31 to 60 minutes
c More than 60 minutes
c Don't Know

B-29 In a **typical week** in the past **3 months**, how much time did you usually spend on a computer, including playing computer games and using the Internet or World Wide Web? **(Please do not include time spent at work or at school)**

c None
c Less than 1 hour
c From 1 to 2 hours
c From 3 to 5 hours
c From 6 to 10 hours
c From 11 to 14 hours
c From 15 to 20 hours
c More than 20 hours

B-30 In a **typical week** in the past **3 months**, how much time did you usually spend watching television or videos?

c None
c Less than 1 hour
c From 1 to 2 hours
c From 3 to 5 hours
c From 6 to 10 hours
c From 11 to 14 hours
c From 15 to 20 hours
c More than 20 hours

EARLY LIFE EXPOSURES

B-31 Have you ever lived on a farm?

c Yes

c No

c Don't know

B-32 Did you live on a farm during your first year of life?

c Yes → **If yes, what type of farm? (Check all that apply)**

c Grain
c Livestock

c No

c Don't know

B-33 Did your mother smoke while she was pregnant with you?

c Yes

c No

c Don't know

B-34 What was your birth weight?
________ pounds or ________ grams

c Don't know

B-35 Were you breastfed as a child?

c Yes → **If yes, was it for 6 months or longer?** c Yes c No

c No

c Don't know

CIGARETTE SMOKING

B-36 Have you ever smoked cigarettes? **(If you have smoked less than 20 packs of cigarettes in your lifetime, answer no.)**

c Yes

c No → **If no, go to question B-43**

B-37 Do you now smoke cigarettes?

c Yes

c No

B-38 How old were you when you first started regular cigarette smoking? ____ years old

B-39 How many cigarettes do you smoke per day now?_____ cigarettes per day

B-40 On the average of the entire time you smoked, how many cigarettes did you smoke per day?
_____ cigarettes per day

B-41 If you have stopped smoking cigarettes completely, how old were you when you stopped?
_____ age stopped

B-42 If there have been periods when you abstained from smoking, indicate total years of abstinence from smoking. _____ years

B-43 Have you ever smoked a pipe regularly? **(Yes means more than 12 oz of tobacco in a lifetime)**

c Yes

c No

B-44 Have you ever smoked cigars regularly? **(Yes means more than 1 cigar a week for a year)**

c Yes

c No

B-45 Do you smoke a pipe or cigars regularly at present?

c Yes

c No

ALCOHOL CONSUMPTION

B-46 During the past 12 months, how often did you drink alcoholic beverages?

c Never
c Less than once a month
c Once a month
c 2 to 3 times a month
c Once a week
c 2 to 3 times a week
c 4 to 6 times a week
c Every day

B-47 How often in the past 12 months have you had 5 or more drinks on one occasion?

c Never
c Less than once a month
c Once a month
c 2 to 3 times a month
c Once a week
c More than once a week

MEDICAL HISTORY

B-48 In general would you say your health is:

c Excellent
c Very Good
c Good
c Fair
c Poor

B-49 During the past 12 months, were you seen by a doctor or other primary care giver for:

	Yes	**No**	**Don't know**
Stomach acidity or reflux?	c	c	c
An ear infection?	c	c	c
An injury?	c	c	c

B-50 Has a doctor or primary care giver ever said you have:

	Yes	**No**	**Don't Know**
Diabetes	c	c	c
Heart Disease	c	c	c
Heart Attack	c	c	c
Hardening of the arteries	c	c	c
High Blood Pressure	c	c	c
Cystic Fibrosis	c	c	c
Tuberculosis	c	c	c
Stroke	c	c	c
Cancer	c	c	c

If yes to cancer, please specify type(s):

____

____

CHEST ILLNESSES

B-51 Has a doctor ever said you had any of the following chest illnesses:

	Chest Illness	During the Past 12 Months		Ever In Your Life	
a.	Attack of bronchitis	c Yes	c No	c Yes	c No
b.	Pneumonia	c Yes	c No	c Yes	c No
c.	Hay Fever	c Yes	c No	c Yes	c No
d.	Sinus Trouble	c Yes	c No	c Yes	c No
e.	Chronic Bronchitis	c Yes	c No	c Yes	c No
f.	Emphysema	c Yes	c No	c Yes	c No
g.	COPD (Chronic Obstructive Pulmonary Disease)	c Yes	c No	c Yes	c No
h.	Sleep Apnea	c Yes	c No	c Yes	c No
i.	Other Chest Illness (Example chest operation) please specify:	c Yes ____________________	c No	c Yes ____________________	c No

B-52 **If yes to Chronic Obstructive Pulmonary Disease (COPD) in question B-51g,** how many times have you required services for COPD from the following places during the **past 12 months**?

Hospital inpatient: ________ times

Emergency room outpatient: ________ times

Doctor's office: ________ times

REST AND SLEEP

B-53 Do you snore?

c Yes

c No → **If no, go to question B-55.**

c Don't know

B-54 If you snore, is your snoring:

c Slightly louder than breathing?

c As loud as talking?

c Louder than talking?

c Very loud - can be heard in adjacent rooms?

B-55 How likely are you to doze off or fall asleep in the situations described below, in contrast to just feeling tired? This refers to your usual way of life in recent times. Even if you haven't done some of these things recently, try to work out how they would have affected you. **Please check one response choice for each situation.**

SITUATION	RESPONSE CHOICES			
	Would never doze	Slight chance of dozing	Moderate chance of dozing	High chance of dozing
Sitting and reading	c	c	c	c
Watching TV	c	c	c	c
Sitting inactive in a public place (e.g., a theatre or a meeting)	c	c	c	c
As a passenger in a car for an hour without a break	c	c	c	c
Lying down to rest in the afternoon when circumstances permit	c	c	c	c
Sitting and talking to someone	c	c	c	c
Sitting quietly after lunch without alcohol	c	c	c	c
In a car, while stopped for a few minutes in the traffic	c	c	c	c

Adult 1

FAMILY HISTORY

B-56 Have the following members of your biological family ever had:

	FATHER			MOTHER			BROTHER/SISTER		
	Yes	No	Don't Know	Yes	No	Don't Know	Yes	No	Don't Know
Diabetes	c	c	c	c	c	c	c	c	c
Heart Disease	c	c	c	c	c	c	c	c	c
Heart Attack	c	c	c	c	c	c	c	c	c
Hardening of the arteries	c	c	c	c	c	c	c	c	c
High Blood Pressure	c	c	c	c	c	c	c	c	c
Cystic Fibrosis	c	c	c	c	c	c	c	c	c
Tuberculosis	c	c	c	c	c	c	c	c	c
Stroke	c	c	c	c	c	c	c	c	c
Lung Trouble (Asthma,Emphysema, Chronic Bronchitis)	c	c	c	c	c	c	c	c	c
Cancer	c	c	c	c	c	c	c	c	c
If yes to cancer, please specify type(s):	____________			____________			____________		

OCCUPATIONAL HISTORY

B-57 Please list all full-time jobs at which you have worked for at least one year, starting with your present or most recent job. Please state the job title and business as specifically as possible. For example, 'mixed farming' instead of 'farming'.

Job Title	Business, Industry or Service	Total number of Years at job
e.g. Nurse	Health Care	10
e.g. Farmer	Mixed Farming	30

B-58 Have you ever been exposed to any of the following in the work place?

	No	Yes	If Yes, how often?		How many years?
Grain Dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Mine dust (e.g. potash, uranium)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Specify ________________					
Asbestos dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Wood dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Other dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Specify ________________					
Livestock	c	c	Daily c Weekly c	Monthly c Occasionally c	
Smoke from stubble burning	c	c	Daily c Weekly c	Monthly c Occasionally c	
Diesel fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Welding fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Solvent fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Oil / Gas well fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Herbicides (to kill plants)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Fungicides (to treat grain)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Insecticides (to kill insects)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Molds	c	c	Daily c Weekly c	Monthly c Occasionally c	
Radiation	c	c	Daily c Weekly c	Monthly c Occasionally c	
Other, Specify ________________	c	c	Daily c Weekly c	Monthly c Occasionally c	

B-59 How often do you (did you) wear a dust mask when exposed to grain dust?

c Always c Most of the time c Sometimes c Never

B-60 We wish to find out more about respiratory health of rural people. Would you be willing to be contacted about having breathing and/or allergy tests at a nearby location?

c Yes

c No

c I would like more information

IF THERE IS ONLY ONE ADULT IN YOUR FAMILY, PLEASE SKIP TO THE LAST PAGE.

IF THERE IS ANOTHER ADULT IN YOUR FAMILY, PLEASE CONTINUE ON THE NEXT PAGE.

REMEMBER TO COMPLETE THE CONTACT INFORMATION ON THE LAST PAGE!
(THIS INFORMATION WILL BE REMOVED FROM YOUR QUESTIONNAIRE TO ENSURE CONFIDENTIALITY.)

SECTION B INDIVIDUAL QUESTIONS

ADULT 2

NOW, PLEASE ANSWER THE FOLLOWING QUESTIONS ABOUT ADULT # 2.

B-1 Age as of January 1st, 2010: ___________

B-2 Date of birth: MM______ DD____ YY_______

B-3 Sex: Male c Female c

B-4 Highest level of education:

c Less than high school
c Completed high school
c Completed university
c Completed post-secondary education other than above

B-5 What is your ethnic background?

c Caucasian
c First Nation
c Metis
c Other → **Please specify:** _____________

B-6 What is your height? _____cm. **OR** ______ft and in.

B-7 What is your weight? ________Kg. **OR** _______lbs

B-8 What is your marital status? **(Please check only one)**

c Married
c Common law/living together
c Widowed
c Divorced/separated
c Single, never married

Adult 2

RESPIRATORY HEALTH

<u>**COUGH**</u>

B-9 Do you usually have a cough?

c Yes
c No → **If no, go to question B-12.**

B-10 Do you usually cough like this on most days for 3 consecutive months or more during the year?

c Yes

c No

B-11 For how many years have you had this cough?
_____ years

PHLEGM

B-12 Do you usually bring up phlegm from your chest? c Yes

c No → **If no, go to question B-15.**

B-13 Do you bring up phlegm like this on most days for 3 consecutive months or more during the year?

c Yes

c No

B-14 For how many years have you had trouble with phlegm?
_____ years

WHEEZE

B-15 Does your chest ever sound wheezy or whistling:

	Yes	**No**
1. When you have a cold?	c	c
2. Apart from colds?	c	c
3. Most days or nights?	c	c

If YES to 1, 2, OR 3, for how many years has this been present? _______________number of years

B-16 Have you ever had an attack of wheezing that has made you feel short of breath?

c Yes

c No

If YES, have you ever required medicine or treatment for the(se) attack(s)?

c Yes

c No

BREATHLESSNESS

B-17 Are you troubled by shortness of breath when hurrying on the level or walking up a slight hill?

c Yes

c No

B-18 Do you have to walk slower than people of your age because of breathlessness?

c Yes

c No

B-19 Do you ever have to stop for breath when walking at your own pace on the level?

c Yes

c No

B-20 Do you ever have to stop for breath after walking about 100 yards (or after a few minutes) on the level?

c Yes

c No

B-21 Are you too breathless to leave the house or breathless on dressing or undressing?

c Yes

c No

ASTHMA

B-22 Have you ever had asthma? c
Yes

c No → **If no, go to question B-26.**

B-23 **If Yes to B-22:**

Do you still have it?	c Yes	c No
Was it confirmed by a doctor?	c Yes	c No

At what age did it start? ____ age in years

If you no longer have it, at what age did it stop?
____ age in years

B-24 **If yes to B-22**, how many times have you required services for asthma from the following places during the past 12 months?

Hospital inpatient: _________ times
Emergency room outpatient: _________ times
Doctor's office: _________ times

B-25 **If yes to B-22,** which of the following statements best describes your asthma medication use in the past 12 months:

c Never in the past 12 months
c At least once in the past 12 months
c At least once per month
c At least once per week
c Every day

ALLERGIES

B-26 Have you ever had an allergic reaction to any of the following: **(Please check all that apply).**

1. House dust	c Yes	c No
2. Cats	c Yes	c No
3. Dogs	c Yes	c No
4. Grasses	c Yes	c No
5. Pollens	c Yes	c No
6. Molds	c Yes	c No
7. Others,	c Yes	c No

Please specify:____________________

PHYSICAL ACTIVITY

B-27 Do you exercise?

c Yes → **If yes, how many times a week?**
_____times a week

c No → **If no, go to question B-29.**

B-28 How long do you usually exercise? c
Less than 15 minutes
c 15 to 30 minutes c
31 to 60 minutes
c More than 60 minutes c
Don't Know

B-29 In a **typical week** in the past **3 months**, how much time did you usually spend on a computer, including playing computer games and using the Internet or World Wide Web? **(Please do not include time spent at work or at school)**

c None
c Less than 1 hour
c From 1 to 2 hours
c From 3 to 5 hours
c From 6 to 10 hours
c From 11 to 14 hours
c From 15 to 20 hours
c More than 20 hours

B-30 In a **typical week** in the past **3 months**, how much time did you usually spend watching television or videos?

c None
c Less than 1 hour
c From 1 to 2 hours
c From 3 to 5 hours
c From 6 to 10 hours
c From 11 to 14 hours
c From 15 to 20 hours
c More than 20 hours

EARLY LIFE EXPOSURES

B-31 Have you ever lived on a farm?

c Yes

c No

c Don't know

B-32 Did you live on a farm during your first year of life?

c Yes → **If yes, what type of farm? (Check all that apply)**

c Grain
c Livestock

c No

c Don't know

B-33 Did your mother smoke while she was pregnant with you?

c Yes

c No

c Don't know

B-34 What was your birth weight?
________ pounds or ________ grams

c Don't know

B-35 Were you breastfed as a child?

c Yes → **If yes, was it for 6 months or longer?** c Yes c No

c No

c Don't know

CIGARETTE SMOKING

B-36 Have you ever smoked cigarettes? **(If you have smoked less than 20 packs of cigarettes in your lifetime, answer no.)**

c Yes

c No → **If no, go to question B-43**

B-37 Do you now smoke cigarettes?

c Yes

c No

B-38 How old were you when you first started regular cigarette smoking? ____ years old

B-39 How many cigarettes do you smoke per day now?_____ cigarettes per day

B-40 On the average of the entire time you smoked, how many cigarettes did you smoke per day?
_____ cigarettes per day

B-41 If you have stopped smoking cigarettes completely, how old were you when you stopped?
_____ age stopped

B-42 If there have been periods when you abstained from smoking, indicate total years of abstinence from smoking. _____ years

B-43 Have you ever smoked a pipe regularly? **(Yes means more than 12 oz of tobacco in a lifetime)**

c Yes

c No

B-44 Have you ever smoked cigars regularly? **(Yes means more than 1 cigar a week for a year)**

c Yes

c No

B-45 Do you smoke a pipe or cigars regularly at present?

c Yes

c No

ALCOHOL CONSUMPTION

B-46 During the past 12 months, how often did you drink alcoholic beverages?

c Never
c Less than once a month
c Once a month
c 2 to 3 times a month
c Once a week
c 2 to 3 times a week
c 4 to 6 times a week
c Every day

B-47 How often in the past 12 months have you had 5 or more drinks on one occasion?

c Never
c Less than once a month
c Once a month
c 2 to 3 times a month
c Once a week
c More than once a week

MEDICAL HISTORY

B-48 In general would you say your health is:

c Excellent
c Very Good
c Good
c Fair
c Poor

B-49 During the past 12 months, were you seen by a doctor or other primary care giver for:

	Yes	**No**	**Don't know**
Stomach acidity or reflux?	c	c	c
An ear infection?	c	c	c
An injury?	c	c	c

B-50 Has a doctor or primary care giver ever said you have:

	Yes	**No**	**Don't Know**
Diabetes	c	c	c
Heart Disease	c	c	c
Heart Attack	c	c	c
Hardening of the arteries	c	c	c
High Blood Pressure	c	c	c
Cystic Fibrosis	c	c	c
Tuberculosis	c	c	c
Stroke	c	c	c
Cancer	c	c	c

If yes to cancer, please specify type(s):

____

____

CHEST ILLNESSES

B-51 Has a doctor ever said you had any of the following chest illnesses:

	Chest Illness	During the Past 12 Months		Ever In Your Life	
a.	Attack of bronchitis	c Yes	c No	c Yes	c No
b.	Pneumonia	c Yes	c No	c Yes	c No
c.	Hay Fever	c Yes	c No	c Yes	c No
d.	Sinus Trouble	c Yes	c No	c Yes	c No
e.	Chronic Bronchitis	c Yes	c No	c Yes	c No
f.	Emphysema	c Yes	c No	c Yes	c No
g.	COPD (Chronic Obstructive Pulmonary Disease)	c Yes	c No	c Yes	c No
h.	Sleep Apnea	c Yes	c No	c Yes	c No
i.	Other Chest Illness (Example chest operation) please specify:	c Yes ________________	c No	c Yes ________________	c No

B-52 **If yes to Chronic Obstructive Pulmonary Disease (COPD) in question B-51g,** how many times have you required services for COPD from the following places during the **past 12 months**?

Hospital inpatient: ________ times

Emergency room outpatient: ________ times

Doctor's office: ________ times

REST AND SLEEP

B-53 Do you snore?

c Yes
c No → **If no, go to question B-55.**
c Don't know

B-54 If you snore, is your snoring:

c Slightly louder than breathing?
c As loud as talking?
c Louder than talking?
c Very loud - can be heard in adjacent rooms?

B-55 How likely are you to doze off or fall asleep in the situations described below, in contrast to just feeling tired? This refers to your usual way of life in recent times. Even if you haven't done some of these things recently, try to work out how they would have affected you. **Please check one response choice for each situation.**

	RESPONSE CHOICES			
SITUATION	Would never doze	Slight chance of dozing	Moderate chance of dozing	High chance of dozing
Sitting and reading	c	c	c	c
Watching TV	c	c	c	c
Sitting inactive in a public place (e.g., a theatre or a meeting)	c	c	c	c
As a passenger in a car for an hour without a break	c	c	c	c
Lying down to rest in the afternoon when circumstances permit	c	c	c	c
Sitting and talking to someone	c	c	c	c
Sitting quietly after lunch without alcohol	c	c	c	c
In a car, while stopped for a few minutes in the traffic	c	c	c	c

FAMILY HISTORY

B-56 Have the following members of your biological family ever had:

	FATHER			MOTHER			BROTHER/SISTER		
	Yes	No	Don't Know	Yes	No	Don't Know	Yes	No	Don't Know
Diabetes	c	c	c	c	c	c	c	c	c
Heart Disease	c	c	c	c	c	c	c	c	c
Heart Attack	c	c	c	c	c	c	c	c	c
Hardening of the arteries	c	c	c	c	c	c	c	c	c
High Blood Pressure	c	c	c	c	c	c	c	c	c
Cystic Fibrosis	c	c	c	c	c	c	c	c	c
Tuberculosis	c	c	c	c	c	c	c	c	c
Stroke	c	c	c	c	c	c	c	c	c
Lung Trouble (Asthma,Emphysema, Chronic Bronchitis)	c	c	c	c	c	c	c	c	c
Cancer	c	c	c	c	c	c	c	c	c
If yes to cancer, please specify type(s):	____________			____________			____________		

OCCUPATIONAL HISTORY

B-57 Please list all full-time jobs at which you have worked for at least one year, starting with your present or most recent job. Please state the job title and business as specifically as possible. For example, 'mixed farming' instead of 'farming'.

Job Title	Business, Industry or Service	Total number of Years at job
e.g. Nurse	Health Care	10
e.g. Farmer	Mixed Farming	30

B-58 Have you ever been exposed to any of the following in the work place?

	No	Yes	If Yes, how often?		How many years?
Grain Dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Mine dust (e.g. potash, uranium)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Specify ________________					
Asbestos dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Wood dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Other dust	c	c	Daily c Weekly c	Monthly c Occasionally c	
Specify ________________					
Livestock	c	c	Daily c Weekly c	Monthly c Occasionally c	
Smoke from stubble burning	c	c	Daily c Weekly c	Monthly c Occasionally c	
Diesel fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Welding fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Solvent fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Oil / Gas well fumes	c	c	Daily c Weekly c	Monthly c Occasionally c	
Herbicides (to kill plants)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Fungicides (to treat grain)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Insecticides (to kill insects)	c	c	Daily c Weekly c	Monthly c Occasionally c	
Molds	c	c	Daily c Weekly c	Monthly c Occasionally c	
Radiation	c	c	Daily c Weekly c	Monthly c Occasionally c	
Other, Specify ________________	c	c	Daily c Weekly c	Monthly c Occasionally c	

How often do you (did you) wear a dust mask when exposed to grain dust?

B-59

c Always c Most of the time c Sometimes c Never

B-60 We wish to find out more about respiratory health of rural people. Would you be willing to be contacted about having breathing and/or allergy tests at a nearby location?

c Yes

c No

c I would like more information

IF THERE ARE MORE THAN TWO ADULT FAMILY MEMBERS LIVING IN YOUR HOUSEHOLD,

PLEASE CONTINUE IN THE GREEN BOOKLET.

REMEMBER TO COMPLETE THE CONTACT INFORMATION ON THE LAST PAGE!
(THIS INFORMATION WILL BE REMOVED FROM YOUR QUESTIONNAIRE TO ENSURE CONFIDENTIALITY.)

CONTACT INFORMATION (PLEASE PRINT)

NAME:______________________________________ Age: _____ c Male c Female

(Name of person completing the survey)

____

Address (number and street and Box Number)

_______________ , _______________

Town Postal code

If you live on a farm give the land location of your residence.

____

Land location (quarter, section, township, meridian)

Telephone Numbers **(check most preferred):**

Work _________________ c

Home _________________ c

Cell _________________ c

THIS IS THE END OF THE SURVEY.

THANK YOU VERY MUCH FOR YOUR HELP!

Appendix B – GLMs with random and systematic components with link functions

Model	Random component	Systematic component	Link function
Analysis of variance (AVOVA)	Normal	Identity	Categorical
Analysis of covariance (ANCOVA)	Normal	Identity	Mixed
Linear regression	Normal	Identity	Continuous
Logistic regression	Binomial	Logit	Mixed
Log-linear regression	Poisson	Log	Categorical
Poisson regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized logit	Mixed

Adopted and modified from Agresti [44].

Appendix C – Notations of Vectors and Matrices

Suppose there are m $(i = 1,2,\dots,m)$ subjects and n_i $(j = 1,2,\dots,n_i)$ measurements on the i^{th} subject taken at times $t_{i1} < t_{i2} < \cdots < t_{in_i}$ in a longitudinal study. If Y_{ij} is the observed responses for the i^{th} subject measured at the t_{ij}, then

$\begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in_i} \end{bmatrix}$ is the $n_i\mathrm{x}1$ column vector of n_i responses for the i^{th} subject, and

$\begin{bmatrix} y_{i1} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{1n_2} \\ y_{m1} \\ y_{m2} \\ \vdots \\ y_{mn_m} \end{bmatrix}$ is the $(\sum_{i=1}^{m} n_i)\mathrm{x}1$ column vector of responses for all the m subjects.

In addition, if $X_{ij1}, X_{ij2}, \dots, X_{ijp}$ are specific observations measured on the set of p covariates $(X_1, X_2, \dots, X_p)$ at time t_{ij} for the i^{th} subject, then the j^{th} measurement associated with the i^{th} subject at time t_{ij} is given by;

$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}$ is a $p\mathrm{x}1$ column vector of covariates and

the design matrix

$X_i = \begin{bmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{in_i} \end{bmatrix} = \begin{bmatrix} X_{i11}, X_{i12}, \dots, X_{i1p} \\ X_{i21}, X_{i22}, \dots, X_{i1p} \\ \vdots \\ X_{in_i1}, X_{in_i2}, \dots, X_{in_ip} \end{bmatrix}$ is an $n_i\mathrm{x}p$ matrix of covariates (for the i^{th} subject) whose row entries represent the covariate measurements at the different times t_{ij} and the column entries represent different covariates.

$\varepsilon_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix}$ is an $n_i\mathrm{x}1$ vector of random errors associated with the i^{th} subject.

Since the n_i outcome measurements $(Y_{i1}, Y_{i2}, \dots, Y_{in_i})$ for the i^{th} subject tend to be correlated in a longitudinal study, we account for all the possible intra-class/within-subject correlations using the following notations. Let R_i be the correlation structure for outcome measurements $(Y_{i1}, Y_{i2}, \dots, Y_{in_i})$ for the i^{th} subject, then

$$R_i = \begin{bmatrix} 1 & Corr(Y_{i1}, Y_{i2}) & \cdots & Corr(Y_{i1}, Y_{in_i}) \\ Corr(Y_{i2}, Y_{i1}) & 1 & \cdots & Corr(Y_{i2}, Y_{in_i}) \\ \vdots & \vdots & \vdots & \vdots \\ Corr(Y_{in_i}, Y_{i1}) & Corr(Y_{in_i}, Y_{i2}) & \cdots & 1 \end{bmatrix}$$

is an n_ixn_i symmetric matrix of within-subject correlations resulting in $n_i(n_i - 1)/2$ possible pairwise correlations for each unique pair of outcome measurements. In particular, the j^{th} row and k^{th} column entry of R_i corresponds to $Corr(Y_{ij}, Y_{ik})$. In longitudinal data analysis, one should necessarily take into consideration, the appropriate structure of R_i for each of the m subjects (and the variations between them). This, therefore, requires us to characterize the variance-covariance matrix Σ_i of the outcome measurements $(Y_{i1}, Y_{i2}, \dots, Y_{in_i})$ for the i^{th} subject. By definition, $Corr\left(Y_{ij}, Y_{ik}\right) = Cov(Y_{ij}, Y_{ik}) / \left\{\sqrt{Var\left(Y_{ij}\right)(Y_{ik})}\right\}$. So R_i is related to Σ_i through $\Sigma_i = {V_i}^{1/2} \mathrm{R}_i {V_i}^{1/2}$ where V_i is defined as;

$${V_i}^{1/2} = \begin{bmatrix} \sqrt{Var(Y_{i1})} & 0 & \cdots & 0 \\ 0 & \sqrt{Var(Y_{i2})} & \cdots & 0 \\ & & \ddots & \vdots \\ \vdots & \vdots & & \\ 0 & 0 & \cdots & \sqrt{Var\left(Y_{in_i}\right)} \end{bmatrix} = Diag[\sqrt{Var(Y_{i1})}, \sqrt{Var(Y_{i2})}, \dots, \sqrt{Var(Y_{i1})}],$$

and we write

$$\Sigma_i = \begin{bmatrix} Var(Y_{i1}) & Cov(Y_{i1}, Y_{i2}) & & \cdots & Cov(Y_{i1}, Y_{in_i}) \\ Cov(Y_{i2}, Y_{i1}) & Var(Y_{i2}) & & \cdots & Cov(Y_{i2}, Y_{in_i}) \\ \vdots & \vdots & \ddots & & \vdots \\ Cov(Y_{in_i}, Y_{i1}) & Cov(Y_{in_i}, Y_{i2}) & \cdots & & Var(Y_{in_i}) \end{bmatrix}.$$

In longitudinal data analysis, there are several covariance structures to choose from. Five (5) commonly available covariance structures in software packages are; (i) exchangeable, (ii) independence, (iii) unstructured, (iv) autoregressive first order [AR(1)] assuming equally space sample units, and (v) autoregressive first order assuming unequally space sample units. For the purpose of this thesis, we only focus on the exchangeable type because we assumed that the correlation between any two outcome measurements on the i^{th} subject is the same. However, I produced the variance-covariance matrices of the remaining four for convenience.

1. Exchangeable variance-covariance (also known as compound symmetry or sphericity): when the correlation between two (2) repeated outcome measurements are the same. Time ordering is irrelevant in this case.

$$R_i = \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & \rho & \cdots & \rho \\ \rho & \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \rho & \cdots & 1 \end{bmatrix}$$

2. Independence variance-covariance: repeated outcome measurements at different times are not correlated.

$$R_i = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

3. Unstructured variance-covariance (aka unspecified.): correlation between repeated outcome measurements at different times are unknown and are estimated separately.

$$R_i = \begin{bmatrix} 1 & \rho_{1,2} & \rho_{1,3} & \cdots & \rho_{1,n} \\ \rho_{2,1} & 1 & \rho_{2,3} & \cdots & \rho_{2,n} \\ \rho_{3,1} & \rho_{3,2} & 1 & \cdots & \rho_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho_{n,1} & \rho_{n,2} & \rho_{n,3} & \cdots & 1 \end{bmatrix}$$

4. Autoregressive first order [AR(1)]: decreasing correlation for farther time points

   (i) When the follow-up time points $(t_1, t_2, \dots, t_n)$ are equal spaced (e.g. one-year intervals). i.e. $|t_{i+1} - t_i| = |t_{i+2} - t_{i+1}|, \forall i = 1,2, \dots, n$

$$R_i = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{n-1} \\ \rho & 1 & \rho & \cdots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \cdots & \rho^{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \cdots & 1 \end{bmatrix}$$

   (ii) When the follow-up time points are not equally spaced.

   i.e. $|t_{i+1} - t_i| \neq |t_{i+2} - t_{i+1}|$

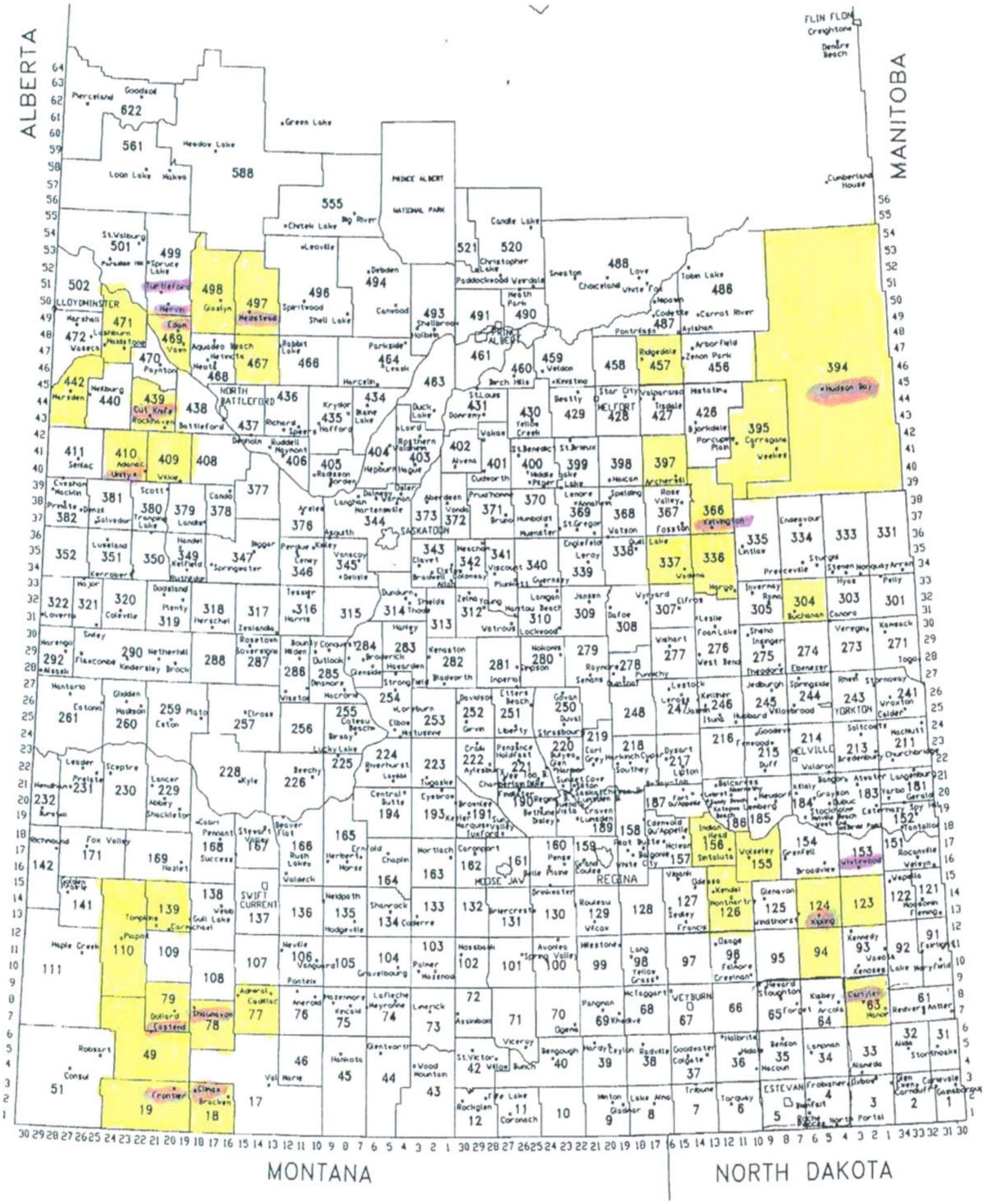
$$R_i = \begin{bmatrix} 1 & \rho^{|t_1-t_2|} & \rho^{|t_1-t_3|} & \cdots & \rho^{|t_1-t_n|} \\ \rho^{|t_2-t_1|} & 1 & \rho^{|t_2-t_3|} & \cdots & \rho^{|t_2-t_n|} \\ \rho^{|t_3-t_1|} & \rho^{|t_3-t_2|} & 1 & \cdots & \rho^{|t_3-t_n|} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{|t_n-t_1|} & \rho^{|t_n-t_2|} & \rho^{|t_n-t_3|} & \cdots & 1 \end{bmatrix}$$

Appendix D – Computational definitions when T is continuous or discrete

Definition/Function	**T is continuous and takes only non-negative values**
Survival function	$S(t) = \mathcal{P}r(T \geq t)$ where $S(0) = 1$ and $S(\infty) = 0$
Probability density function (pdf)	$f(t) = \lim\limits_{\Delta t \to 0} \mathcal{P}r(t \leq T \leq t + \Delta t)/\Delta t = -S'(t) = F'(t)$
Cumulative density function (cdf)	$F(t) = \mathcal{P}r(T < t) = 1 - S(t)$
Hazard function	$h(t) = f(t)/S(t), t \geq 0$
Cumulative hazard function	$H(t) = \int_0^t h(\theta)\, d\theta$

Definition/Function	**T is discrete and takes values $\boldsymbol{t_1, t_2}, \ldots$ such that $\boldsymbol{t_i < t_{i+1}}$**
Survival function	$f(t) = \sum f_i I(t_i < t)$ where $I(\tau) = \begin{cases} 1, & \text{if } \tau \text{ is true} \\ 0, & \text{otherwise} \end{cases}$
Probability density function (pdf)	$f(t) = \begin{cases} f_i & if\ t = t_i \\ 0 & if\ t \neq t_i \end{cases}$ where $f_i = \mathcal{P}r(T = t_i)$
Cumulative density function (cdf)	$F(t) = \sum f_i I(t_i < t) = 1 - S(t)$
Hazard function	$h(t) = f(t)/S(t), t \geq 0$
Cumulative hazard function	$H(t) = \sum h(t_i) I(t_i \leq t)$

Appendix E – Rural municipalities located in the four study quadrants of the SRHS Study.


ALBERTA
MANITOBA
MONTANA
NORTH DAKOTA
FLIN FLON
Creighton
Cumberland House
PRINCE ALBERT NATIONAL PARK
LLOYDMINSTER
NORTH BATTLEFORD
PRINCE ALBERT
MELFORT
SASKATOON
YORKTON
MELVILLE
MOOSE JAW
REGINA
SWIFT CURRENT
WEYBURN
ESTEVAN
622
561
588
555
501
499
502
498
497
496
494
493
491
490
521
520
488
487
486
472
471
470
469
468
467
466
464
463
461
460
459
458
457
456
442
440
439
438
437
436
435
434
431
430
429
428
427
426
411
410
409
408
406
405
404
403
402
401
400
399
398
397
395
394
382
381
380
379
378
377
376
373
372
371
370
369
368
367
366
352
351
350
349
347
346
345
344
343
342
341
340
339
338
337
336
335
334
333
331
322
321
320
319
318
317
316
315
314
313
312
310
309
308
307
305
304
303
301
292
290
288
287
286
285
284
283
282
281
280
279
278
277
276
275
274
273
271
261
260
259
257
256
255
254
253
252
251
250
248
247
246
245
244
243
241
232
231
230
229
228
226
225
224
223
222
221
220
219
218
217
216
215
214
213
211
194
193
191
190
189
188
187
186
185
184
183
181
171
169
168
167
166
165
164
163
162
161
160
159
158
157
156
155
154
153
152
151
142
141
139
138
137
136
135
134
133
132
131
130
129
128
127
126
125
124
123
122
121
111
110
109
108
107
106
105
104
103
102
101
100
99
98
97
96
95
94
93
92
91
79
78
77
76
75
74
73
72
71
70
69
68
67
66
65
64
63
61
51
49
46
45
44
43
42
40
39
38
37
36
35
34
33
32
31
19
18
17
12
11
10
9
8
7
6
5
4
3
2
1