Completeness of Rheumatoid Arthritis Prevalence Estimates from Administrative Health Data: Comparison of Capture-Recapture Models

by

Yao Nie

A Thesis submitted to the Faculty of Graduate Studies

The University of Manitoba

In partial fulfillment of the requirement of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

Faculty of Medicine

University of Manitoba

Winnipeg

Copyright © 2014 by Yao Nie

ABSTRACT

Rheumatoid arthritis (RA) is a chronic disease characterized by an overactive immune system and joint inflammation. Population-based administrative health data (AHD) are widely used for RA outcomes research and surveillance. However, AHD may not completely capture all cases of RA in the population. Capture-recapture (CR) methods have been proposed to describe the completeness of AHD for estimating disease population size, but AHD may not conform to the assumptions that underlie CR models. A Monte Carlo simulation study was used to investigate the effects of violations of the assumptions for two-source CR methods: dependence between data sources and heterogeneity of capture probabilities. We compared the Chapman estimator and an estimator based on the multinomial logistic regression model (MLRM) to study relative bias (RB), coverage probability (CP) of 95% confidence intervals, width of 95% confidence intervals (WCI), and root-mean-square-error (RMSE) in prevalence estimates. The effects of misspecification of the MLRM were also investigated. In addition, the Chapman and MLRM estimators were used to estimate RA prevalence using AHD data from Saskatchewan, Canada. Population sizes were consistently underestimated for CR methods when the assumptions were violated. The estimated population size for both of the estimators did not differ substantially except for the RMSE values. Parameter estimates became biased when the MLRM model was misspecified, but there was little impact on population size estimates. In conclusion, CR methods are recommended to reduce bias in prevalence estimates based on AHDS. Because these methods may be sensitive to assumption violations, researchers should consider potential dependence between data sources. As well, sufficient overlap in the cases captured by each data

source (e.g., 50% of the cases are captured by both data sources) or balanced capture probability in each data source is needed to effectively implement these methods. Researchers who estimate population size using CR methods in AHDs should favour the MLRM estimator over the Chapman estimator.

ACKNOWLEDGMENTS

To start, I would like to thank my supervisor, Dr. Lisa Lix, for her consistent guidance and support in my thesis work. We have been building up this work step-by-step and each time she has given feedback so that I can improve. From Saskatoon to Winnipeg, you have been a valuable source of advice, for both of my M.Sc. program and my thesis work, that I am really thankful. Also, I would like to thank Dr. Depeng Jiang who has taught me to think differently towards this work. As well, thank you to my committee members, Dr. Nazeem Muhajarine, and Dr. Natalie Shiff, who have also contributed to my thesis work with their insight input and suggestions.

This work would not be possible without the previous work done by other researchers. The valuable information in your work laid the foundation of my thesis. Starting from the scratch, I really appreciate that I can build up skills in the population data lab with so many wonderful people working around me every day. I am also really thankful for the support that I have received from the Department of Community Health Sciences and students who kept company with me during the lunch time.

I would also like to thank all of my friends in both Saskatoon and Winnipeg. Thank you, to Dr. Clement Yeung and Rosalind Yeung for being a consistent and invaluable source of encouragement during my life in Winnipeg. Finally, to my parents in China for your unconditional love in the past twenty-five years – I love you. You may not totally understand what I am doing for my degree, but you do know that education plays an important role in my life. From my childhood, you have made everything available for me such as learning English which became an advantage to me for studying abroad. Thank you!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AHD	Administrative health data
CCHS	Canadian Community Health Survey
CIHR	Canadian Institutes of Health Research
CIs	Confidence intervals
CP	Coverage probability
CR	Capture-recapture
DM	Dermatomyositis
DMARDs	Disease modifying antirheumatic drugs
GP	General practitioner
ICD	International Classification of Diseases
ICD-9	International Classification of Diseases, 9^{th} revision
ICD-10-CA	International Classification of Diseases, 10^{th} revision, Canada
IML	Interactive Matrix Language
LRT	Likelihood ratio test
MLE	Maximum likelihood estimator
MLRM	Multinomial logistic regression model
NPV	Negative predictive value
PM	Polymyositis
PPV	Positive predictive value
RA	Rheumatoid arthritis
RB	Relative bias
RF	Rheumatoid factor
RMSE	Root-mean-square-error
SARDs	Systemic autoimmune rheumatic diseases
SES	Socioeconomic status
SjS	Sjogren's syndrome
SLE	Systemic lupus erythematosus
SSc	Systemic sclerosis
VA	Veterans administrative
WCI	Width of 95% confidence interval

CHAPTER 1 INTRODUCTION

1.1 Background

Rheumatoid arthritis (RA) is a chronic disease characterized by an overactive immune system and joint inflammation. It is associated with recurrent periods of pain, fatigue, and stiffness, as well as progressive functional disability. It affects approximately one percent of the Canadian population (Toronto Western Research Institute, 2010). The causes of RA are unknown, but risk factors include socioeconomic status (SES), sex, geography, and ethnicity (Barton et al., 2011; Gabriel, 2001; Michaud & Wolfe, 2007; Waltz, Kriegel, & Van't Pad Bosch, 1998).

The Arthritis Alliance of Canada (Bombardier, Hawker, & Mosher, 2011) estimated that more than 272,000 people were living with RA in 2010, which is 0.9% of the Canadian adult population. This number is expected to increase to 549,218 (i.e., 1.3% of the Canadian adult population) over the next 30 years as the population ages. By 2040, the number of new cases of RA in Canada is predicted to be about 23,732, up from 17,916 cases in 2010 (Bombardier et al., 2011).

RA prevalence has been reported to vary internationally (Rasch, Hirsch, Paulose-Ram, & Hochberg, 2003; Toronto Western Research Institute, 2010). Prevalence is estimated to be higher in Australia and lower in Sub-Saharan Africa (Shapira, Agmon-Levin, & Shoenfeld, 2010). In the USA, RA prevalence was estimated to range from 2.03% to 2.72% among respondents 60 years and older based on data from the National Health and Nutrition Examination Survey from 1988 to 1994 (Simard & Mittleman, 2007). Data from the UK's Norfolk Arthritis Registry suggested that the adult population

(i.e., 18+ years) in that region had an estimated RA prevalence of about 0.81% (Symmons et al., 2002).

RA not only impacts the individual, but also the health care system. RA is associated with a number of comorbid conditions such as depression, cardiovascular disease, and cancer, which can adversely affect patients' quality of life and need for health care treatment (Michaud & Wolfe, 2007). Disease modifying antirheumatic drugs (DMARDs) and biologic therapies, which are commonly used to treat patients with RA, are expensive. It is currently estimated that RA drives more than $2 billion in direct health care costs in Canada (Bombardier et al., 2011). If severe RA cases could be avoided, Bombardier et al. (2011) also estimated that $5.1 billion could be saved in cumulative direct health care costs and over $33.7 billion could be saved in cumulative productivity losses over 30 years in Canada.

Given the impact on the individual and the health care system, epidemiologic studies about RA prevalence and incidence are important to provide an indication of both the overall and relative burden of the disease. In Canada, the two main sources of population-based data for RA epidemiologic studies are national survey data from such sources as the Canadian Community Health Survey (CCHS) and administrative health data (AHD). Both sources of data have their strengths and limitations. Given the low prevalence of RA in the population, national survey data, which are based on samples of the population, are likely to produce less precise estimates than AHD. As well, previous research has shown that the CCHS tends to result in substantially over-estimated prevalence of RA (L. Lix et al., 2006), possibly because of the manner in which the

questions about arthritis are worded. As a result, the self-report question in CCHS (e.g., cycle 1.1) was not effective in distinguishing different forms of arthritis, including RA.

AHD also have some limitations for estimating RA prevalence and incidence. For example, hospital data had a low sensitivity of 23% when compared with rheumatologist-reported diagnosis and a single physician visit had a specificity of only 60% for ascertaining disease cases (Widdifield et al., 2013). In addition, sensitivity only ranged from 5.0% to 11.3% for RA algorithms applied to AHD (L. Lix et al., 2006), when using survey data (i.e., CCHS) as the reference standard although the results may be influenced by the choice of a gold standard. Thus, AHD could result in biased estimates of prevalence and incidence, with low sensitivity to detect true positive cases being of particular concern.

Methods to adjust for less-than-perfect sensitivity and specificity of AHD include the use of model-based prediction algorithms (L. M. Lix, Yogendran, Leslie, et al., 2008), bias-corrected prevalence or incidence estimates (Manuel, Rosella, & Stukel, 2010), latent class analysis techniques (Bernatsky et al., 2011), and capture-recapture (CR) methods (Hook & Regal, 1995). None of these methods is without limitations and each make assumptions about the characteristics of the data and the underlying statistical model.

CR methods can be used to describe the completeness of AHD for estimating disease population size. They have been used in a number of epidemiologic studies for this purpose (Yip et al., 1995). However, two assumptions of some CR methods, independence of data sources and homogeneity of capture probabilities, are unlikely to be satisfied in practice. Ascertainment in different AHDs may not be independent; for

example, the probability of being identified as a case in physician billing records may be associated with the probability of being captured in prescription drug records. And this would be a violation of the assumption of independence of data sources. As well, capture probabilities may not be consistent (i.e., homogeneous) for all individuals in the population, and may be associated with such characteristics such as age, sex, and the presence of co-morbid conditions.

1.2 Research Purpose and Objectives

The overall purpose of this research was to apply CR methods to estimate the completeness of AHD for ascertaining RA prevalence. CR methods were chosen for this study because they have not previously, to the best of our knowledge, been applied to estimate RA prevalence, even though they have been applied to other chronic diseases such as cancer (McClish & Penberthy, 2004), diabetes (Giarrizzo, Pezzotti, Silvestri, & Di Lallo, 2007), and stroke (Tilling, Sterne, & Wolfe, 2001) and to AHD in other jurisdictions. In order to know the validity of prevalence estimates from CR methods, we also conducted a Monte Carlo simulation study, to investigate the effect of dependence amongst data sources and heterogeneity of capture probability on estimates of population size.

1.2.1 Research Objective 1: To Compare Population Size Estimates from Conventional and Model-based CR Methods

We compared the completeness of AHD for ascertaining RA prevalence using CR models with and without adjustment for measured covariates. The Chapman estimator and an estimator based on the multinomial logistic regression models (MLRM) were used in both a Monte Carlo simulation study and a numeric example to estimate prevalence.

1.2.2 Research Objective 2: To Explore the Effects of Model Misspecification for the Model-based CR Method

We compared bias in the estimates of completeness of AHD for ascertaining RA prevalence using CR models that were and were not misspecified due to unmeasured covariates. The effects of misspecification were investigated through a Monte Carlo simulation study.

1.3 Thesis Organization

This thesis focuses on epidemiologic methods in chronic disease surveillance. In Chapter 2 we present relevant background on each of the following major topics: 1) Population-based chronic disease prevalence estimation methods, 2) Accuracy of case ascertainment algorithms for chronic disease in AHD, 3) Methods to address bias in prevalence estimates of chronic disease in AHD, 4) CR methods, and 5) Monte Carlo simulation studies about CR methods. Chapter 3 presents the CR methods, describes the dataset to be used for RA prevalence estimates in a numeric example and provides detailed information on the simulation and modelling techniques to compare CR methods. Chapter 4 presents results and analysis from the Monte Carlo simulation study and the numeric example. We finish the thesis with discussion of the key findings and conclusions.

CHAPTER 2 LITERATURE REVIEW

This review provides background information about population-based chronic disease prevalence estimation methods from health survey and AHD. The review then moves on to discuss sensitivity and specificity of AHD for chronic disease case ascertainment. Next, methods to address diagnostic bias and prevalence estimates for chronic disease in AHD are reviewed. CR methods are discussed in detail, including their history, assumptions, and underlying statistical models. Finally, Monte Carlo simulation studies about CR methods are reviewed.

2.1 Population-based Chronic Disease Prevalence Estimation Methods

There are two main population-based data sources used in RA epidemiologic and health outcomes research in Canada: health surveys and AHD. Health surveys require the participants' subjective judgment of their health and recall of past health events. The accuracy of the participants' answers to specific questions may be affected by various factors. Responses may vary according to the method of data collection, the precise phrasing of the questions, and the respondents' understanding of their health and disease (Young, 2005). It is relatively easy to produce health indicators from survey data.

In Canada, for example, the CCHS is a key source of data for RA research and surveillance (Statistics Canada, 2013). The CCHS is conducted by Statistics Canada to provide cross-sectional self-reported information about health determinants, health status, and health system utilization for populations in 133 health regions across Canada (Manitoba Centre for Health Policy, 2007; Statistics Canada, 2009). The CCHS initiative began in 2000 with its main goals being "the provision of population-level information on health determinants, health status and health system utilization" (Health Canada,

2012). There is an annual component on general health; a component about specific health topics is also conducted every two to three years.

AHD are collected by governments for administrative purposes, such as keeping track of the population eligible for health benefits, paying doctors, or funding hospitals. Examples of AHD are hospital abstracts, physician billing claims, and prescription drug records (Spasoff, 1995). Wenneberg and Gittelsohn published one of the first articles using AHD to describe variations in health care use in the United States (Wennberg & Gittelsohn, 1973). Since then, applications of AHD for epidemiologic and health outcomes research have become increasingly common. AHD are cost-effective and time-effective to use for research and surveillance because they are routinely collected. At the same time, AHD can usually be accessed without patient-specific consent, which can reduce selection bias (Suissa & Garbe, 2007). They contain consistent elements, can be accessed in a timely manner, and provide information about large cohorts (Virnig & McBean, 2001).

Reliability about precision of the information found in population-based data have been addressed by many researchers (e.g., Virnig & McBean (2001)). AHD and survey data may not always produce consistent results. Study which compared chronic disease case ascertainment between the CCHS (cycle 1.1) and AHD has been shown (L. M. Lix, Yogendran, Shaw, et al., 2008). Agreement was high for diabetes and hypertension but low for arthritis. For example, algorithms based on only one physician claims contact had a Cohen's kappa coefficient of 0.69 with 95% CIs of (0.68, 0.69) for diabetes and 0.64 with 95% CIs of (0.64, 0.64) for hypertension. However, it was only 0.27 with 95% CIs of (0.26, 0.27) for arthritis. It has been argued that the non-life-threatening nature of

arthritis may result in it being over reported in surveys, but underreported in AHD, contributing to the lack of agreement between the two data sources (Kriegsman, Penninx, Van Eijk, Boeke, & Deeg, 1996). Like arthritis, RA also had a poor agreement between AHD and survey data with a Cohen's kappa coefficient of 0.17 for the algorithm based on one or more AHD (L. Lix et al., 2006). The lack of agreement between survey and administrative data for all forms of arthritis, including RA, may be attributed to the wording of the survey questions, self-report bias, or sampling bias in survey data or due to diagnostic misclassification in AHD (Wunsch, Harrison, & Rowan, 2005).

2.2 Accuracy of Case Ascertainment Algorithms for Chronic Diseases in AHD

The International Classification of Diseases (ICD) developed by the World Health Organization is typically used to assign diagnosis codes in AHD. The accuracy of the diagnostic codes can be assessed by comparing them with a reference standard. Medical records are frequently used as a reference standard for AHD, although self-report survey data and clinical registries can also be used (Virnig & McBean, 2001).

Several studies have examined the accuracy of diagnosis codes for case ascertainment in AHDs for such chronic conditions as diabetes and hypertension. For example, Hebert et al. compared self-reported diabetes from the Medicare Current Beneficiary Survey with diagnoses of diabetes in Medicare administrative data. Using self-reported diabetes status as the reference standard, they found that in order to get adequate sensitivity (>=70%), specificity (97.5%), researchers should combine information from different types of Medicare claim files, use two years of data to identify cases, and require at least two diagnoses of diabetes among claims involving ambulatory care (Hebert et al., 1999).

One study proposed using a logistic regression model to quantify the probability that a person has kidney disease from multiple markers of the disease (Van Walraven et al., 2010). Without available data for a "gold standard", they developed an accurate and well-calibrated multivariable model that demonstrates the probability that a particular patient in an AHD has kidney disease. They found that the sensitivity of a kidney disease diagnostic code for true kidney disease was very low at 38%. However, the specificity was very high (i.e., 98.9%). The study concluded that multiple variables can be combined to quantify the probability that a person has a particular disease. A multivariable model can significantly increase the accuracy of disease identification in AHD.

A few studies have examined the accuracy of ICD codes for RA within AHD. A retrospective chart abstraction study was conducted for a random sample of patients seen in Ontario rheumatology clinics (Widdifield et al., 2013). Using the medical records at each rheumatologist's clinic and charts as the reference standard, these patients were identified in combinations of RA-coded physician billings and primary and secondary hospital discharge diagnoses (using ICD-9 and ICD-10 diagnosis codes 714 and M05-M06 respectively), and prescription drug claims (for glucocorticoids, DMARDs, and biologic agents). Overall, Widdifield et al. (2013) found that physician billing claims had sensitivity ranging from 94% to 100% while hospital records had a sensitivity of only 23%. Specificity and positive predictive value (PPV) were moderate (e.g., 60% and 55% respectively for only one contact in physician billing claims) to excellent (e.g., 96% and 76% respectively for one contact in hospital records) and increased when multiple general practitioner (GP) billing claims or specialist billing claims were used to ascertain cases (e.g., 80% for contacting a specialist within one year). RA prescription drug claims

slightly decreased sensitivity, but also slightly increased specificity and PPV. The addition of hospital data to physician billing claims had little impact on sensitivity or PPV.

The accuracy of an ICD-9 diagnosis for RA (i.e., ICD-9 code 714) was investigated in a Veterans Administrative (VA) hospital database from the USA (Singh, Holmgren, & Noorbaloochi, 2004). Using chart documentation of RA diagnosis by a rheumatologist on at least two visits at least six weeks apart as the reference standard, Singh et al. (2004) found that this diagnosis code had 100% sensitivity, but specificity was only 55%. The addition of a positive laboratory test for rheumatoid factor and a DMARD prescription for a hospital diagnosis significantly improved specificity to between 83% and 97% and PPV increased to between 81% and 97%, although sensitivity dropped to between 76% and 88%.

Using the CCHS as the reference standard, Lix et al. (2006) proposed a number of algorithms for ascertaining cases of RA in AHD. For these algorithms, sensitivity ranged from 5.0% to 11.3%. The highest sensitivity was for a five-year algorithm based on one or more physician billing claims. Specificity was near 100% for all algorithms. The PPV of an RA ranged from 55.9% to 80.6% and the negative predictive value (NPV) was approximately 92%.

2.3 Methods to Estimate Chronic Disease Prevalence in AHD

Several methods have been proposed to estimate chronic disease prevalence in AHD when sensitivity and specificity of disease diagnosis codes are less than perfect. Manuel et al. (2010) proposed adjusting disease prevalence estimates using sensitivity and specificity estimates from validation studies in order to improve the accuracy of

prevalence estimates. They calculated the potential percentage of misclassified cases, that is false positive cases and false negative cases, based on estimates of sensitivity and specific. They focused on the incidence and prevalence of diabetes for Ontario using the Ontario Diabetes Database (Lipscombe & Hux, 2007), which ascertains cases of diabetes using ICD diagnoses in hospital records and physician billing claims. After Manuel et al. (2010) applied validation study estimates of sensitivity and specificity to the Ontario data, the estimated unbiased prevalence of diabetes in 2005 was 7.2%. This number was 19% lower than the estimate in the original study that did not account for sensitivity and specificity when estimating disease prevalence (8.9%).

Bernatsky et al. (2011) proposed using Bayesian latent class models to deal with under-ascertainment of cases of systemic autoimmune rheumatic diseases (SARDs) in AHD. This methodology identifies disease clusters (i.e., disease present/absent) from imperfect markers of disease status and prior information about the sensitivity and specificity of each of the imperfect markers of disease which do not assume the existence of a gold standard. Bayesian methods assume that "unknown values for a parameter have probability distributions". Bernatsky et al. (2011) set prior inputs for the Bayesian model based on the previous research. The total prevalence was between 2 to 3 per 1,000 cases. The highest prevalence was seen among women who were 45 years older. The estimated SARDs prevalence by using the Bayesian latent class models was very close to the existing North American estimates by using other data sources such as population survey (Bernatsky et al., 2009; Helmick et al., 2008; Kabasakal et al., 2006).

Lix et al. (2008) proposed classification models, including logistic regression models and non-parametric classification trees, to develop model-based case

ascertainment methods that use multiple disease indicators to ascertain osteoporosis cases. The logistic regression performs well when data were characterized by linear associations between the disease markers and the probability of being a case. The nonparametric classification methodology is based on recursive partitioning, which forms homogeneous subgroups in the data. Prevalence estimates from the logistic regression models (i.e., 12.35%) were higher than the estimate from the non-parametric classification trees (i.e., 7.61%) among Manitoba women 50+ years. Compared to other population-based studies, the estimated prevalence of osteoporosis was around 12% which suggests that the logistic regression models produced closer osteoporosis prevalence estimates to those found in the literature (Yang et al., 2006).

2.4 Capture-Recapture Methods

2.4.1 History

Capture-recapture (CR) methods have also been proposed to estimate the completeness of AHD for both acute and chronic disease prevalence estimation (McClish & Penberthy, 2004; Peragallo, Urbano, Lista, Sarnicola, & Vecchione, 2011). CR methods have been used in a number of epidemiologic studies to estimate or adjust for incomplete case ascertainment, by using information on the amount of overlaps in lists of cases identified from distinct sources (Hook & Regal, 1995).

CR methods and models were initially applied in biological studies about the size of fish and wildlife populations, in which a sample of wildlife was captured and tagged, then another sample was taken; the re-captured animals were counted and used to estimate total population size. These techniques were later extended to population health research involving record linkage (Yip et al., 1995) for diseases such as cancer

(Schmidtmann, 2008), tuberculosis (van hest et al., 2007), diabetes (Giarrizzo et al., 2007), HIV (Bernillon, Lievre, Pillonel, Laporte, & Costagliola, 2000), and tuberculosis (Tilling et al., 2001), to estimate completeness of various data sources including AHD.

2.4.2 Assumptions

The assumptions underlying conventional CR techniques are that the population is closed (i.e., individuals who migrate or who otherwise do not have complete health insurance coverage are not included), individuals can be uniquely identified, the probability of capture is homogeneous for all individuals in the population, and captures are independent, conditional on the data source (Hook & Regal, 1995). These assumptions may not be satisfied in AHD (Young, 2005), particularly the assumptions of conditional independence of the data sources and homogeneity of capture probabilities. The likelihood of being captured in one data source may increase the likelihood of capture in another data source. For example, severe cases are more likely to be captured by different sources than less severe cases which lead to positive dependence of case ascertainment. Patient characteristics, including socio-demographic variables and measures of disease severity, may be associated with the likelihood of disease case ascertainment in AHD. However, AHD often contain sparse information on variables that may be associated with the likelihood of capture. While socio-demographic variables such as age, sex, and residence location are often available in the data, measures of disease severity are noticeably absent.

2.4.3 Conventional Method

We can subdivide CR methods into two-source and multiple-source methods, based on the number of data sources used to estimate population size. The conventional

method for two sources is based on a two-by-two contingency table. One may estimate either the number of missed cases or the size of the complete population using the method initially proposed by Peterson and Lincoln (Lincoln, 1930; Petersen, 1896), and further refined for the sparse cell size problem by Chapman (Chapman, 1951). In this thesis, we used the Chapman estimator as a representative of the conventional CR method. The Chapman estimator is based on the well-known maximum likelihood estimators which assume independence of ascertainment by both data sources. Peragallo et al. (2011) used the Chapman estimator to estimate the cancer incidence from cancer diagnosis data in military hospitals and unit infirmaries. They concluded that the estimated incidence of cancer by Chapman's estimator was generally lower than expected which may due to the positive dependence between the data sources.

2.4.4 Model-based Approach

To address violations of the assumptions of conventional CR method, a number of statistical models have been proposed (Tilling & Sterne, 1999). One solution that is relatively straight-forward for epidemiologists and surveillance staff to implement is to adopt a CR regression model in which the probability of capture is modeled as a function of covariates that may be associated with heterogeneity of capture probability. The MLRM have been proposed for the two-source CR problems (Alho, 1990). One can relate the characteristics of disease cases to their probability of being captured by each source using the MLRM. However, lack of conditional independence between data sources still remains in applications to AHD by using the MLRM (McClish & Penberthy, 2004).

McClish & Penberthy (2004) proposed CR techniques, including the MLRM estimator applied to hospital discharge and cancer registry administrative data to estimate the missing number of breast, lung, colorectal, and prostate cancer cases. The MLRM incorporated covariates such as demographic variables, whether or not the hospital had a cancer program that was certified by the American College of Surgeons, and whether or not the patient had surgery as initial treatment for his/her cancer. McClish & Penberthy (2004) concluded that the MLRM can improve the estimate of the number of cancer cases, compared to only counting the number of cases from individual data sources. They also found that demographic variables alone did not account for much of the heterogeneity in capture probabilities. However, the MLRM allows multiple covariates to be taken into account simultaneously.

In multiple-source CR methods, log-linear methods can be used (Hook & Regal, 1995). The use of log-linear models (e.g., using Poisson regression) makes two major assumptions about the capture probabilities. First of all, the capture probabilities for different data sources are not all dependent. Secondly, the capture probability of a source is assumed to be homogeneous for each individual in the population (Tilling & Sterne, 1999). The methods also make the same assumptions of multi-source independence of ascertainment of individuals and variable catchability. It may be difficult to satisfy these assumptions. Hook & Regal (1995) suggested that one way to enhance the plausibility of these assumptions was "the use of as many sources and as many qualitatively different types of sources as possible". By specifying dependencies at one level, and then invoking heterogeneity to additional parameters for each of levels might be acceptable to control violations of assumption in log-liner models (Yip et al., 1995). An alternative way for

reducing heterogeneity is to consider stratified analysis as Yip et al. (1995) demonstrated. To be more precise, investigator should consider stratifying the population of interest by known factors such as demographic variables across different strata. However, Tilling & Sterne (1999) proposed that the increase in stratification has an impact on modeling dependence within strata.

2.5 Monte Carlo Simulation Studies about CR Methods

Monte Carlo simulation studies are defined as "random experiments on a computer" (Kroese, Taimre, & Botev, 2013). Monte Carlo methods are computational algorithms that rely on repeated random sampling to obtain numerical results. The methods were created during the Second World War for the development of the atomic bomb, and since then, they are widely used in science, engineering, finance, and statistics.

Several simulation studies had been conducted to evaluate the performance of CR methods (Alho, 1990; Tilling & Sterne, 1999; Wittes, 1972). The measures that have been used to evaluate performance include mean/mean estimated population size, mean standard deviation, and coverage probability. Wittes (1972) suggested that Chapman's estimator for population size in two-source CR method was unbiased when the sum of the sample sizes from the capture and recapture was no less than the unknown population size. He assumed the cases being captured for twice had a hypergeometric distribution. This indicated that the overlapping cases between two data sources were really high. Wittes (1972) concluded that the estimated population size had unacceptably large negative bias when the number of re-captured cases was small.

Alho (1990) introduced the logistic regression algorithms accounting for population heterogeneity in estimating population size. He allowed different capture probabilities across individuals and across capture times. The resulting regression parameter estimates can be used to estimate the proportion of the population missed, assuming that the population has the same covariate distribution as the sample. And he also compared the variance estimates between conventional method and the logistic regression algorithms. He examined both the asymptotic and finite-sample properties of the proposed estimator. And his results suggested that the model can be widely used when required covariate information existed.

Tilling & Sterne (1999) compared the MLRM and the log-linear model estimating population size. They particularly looked into the effects of covariates on the log-linear model and the MLRM. Bootstrap methods were used to derive the variance for the estimate of population size. They concluded that for CR data without covariates or with categorical covariates, the log-linear model was equivalent to the logit model in terms of performance. When there was dependence between the data sources, the estimated population sizes can be seriously biased. Including covariates in CR methods can reduce the bias in estimating population size. However, the distribution of the covariates may not always be the same in the observed and unobserved segments of the population, which may further bias the estimates of population size according to what Alho (1990) proposed.

2.6 Summary

In summary, RA is a chronic inflammatory disease that affects quality of life and health care utilization. Population-based AHD are widely used for RA surveillance, as

well as health outcomes and service utilization research. However, a limitation of these databases is that they may not capture all RA cases in the population (Singh et al., 2004). CR methods represent one approach to estimate completeness of capture. However, the assumptions of CR models, which include independence of data sources and homogeneity of capture probabilities, might not be satisfied in AHD. The two- source conventional CR method and the model-based CR method were proposed in several literatures. However, few studies have simultaneously considered the two main assumptions of CR methods and the effect of model misspecification on prevalence estimates. This study applied CR methods in both simulation and numeric example to evaluate the performance of conventional method and the model-based approach and to examine the effect of model misspecification through the model-based approach CR method.

CHAPTER 3 METHODS

3.1 Two-source CR Models

To achieve the research objectives, we conducted a simulation study about two-source CR methods, given that it is not possible to investigate the potential biasing effects of assumption violations and unmeasured covariates in real-world data. As well, CR models were applied to estimate completeness of AHD for calculating RA prevalence using an existing dataset from Saskatchewan, Canada.

The definition of "sources" in epidemiologic studies for CR methods is different from that used in animal ecology studies, which the latter usually has a natural temporal ordering. In our research, the two-source CR model is applied to two different AHD sources: physician billing claims and hospital discharge abstracts. Accordingly, there are three possible combinations of these sources from which cases can be ascertained: physician billing claims only, hospital discharge abstracts only, and both data sources. The number of cases missed from both sources is estimated.

3.1.1 Chapman Estimator

Table 3.1 shows the structure of the data and the formulas to estimate the number of missed cases by using the conventional two-source CR method. Let S_1 and S_2 be the total number of cases identified by each source, and a be the cases captured by both sources. The unknown number of cases missed by both sources (x) is to be estimated. Using the maximum likelihood estimator (MLE), the probability of being captured by both data sources is the product of the probability of being captured in Source 1 and Source 2. Accordingly,

$$\frac{a}{N} = \left(\frac{S_1}{N}\right)\left(\frac{S_2}{N}\right) = (S_1 S_2)/N^2, \quad (1)$$

and the probability of being missed by both data sources is

$$\frac{x}{N} = x/(a + b + c + x). \quad (2)$$

Table 3.1 AHD structure for two-source CR problem

		Diagnosis in Source 1		
		Yes	No	Total
Diagnosis in Source 2	Yes	a	b	$S_2 = a + b$
	No	c	x	
	Total	$S_1 = a + c$		$N = a + b + c + x$

Estimated values	**Maximum likelihood estimator (MLE)**
Unobserved cell: $\hat{x}$	bc/a
Completeness of Source 1	$a/(a + b)$
Completeness of Source 2	$a/(a + c)$
Total population: $\widehat{N}$	$a + b + c + (bc/a)$

An adjustment was suggested by Chapman (1951) to reduce the effects of small sample bias on the MLE (Hook & Regal, 1995). Using the Chapman method, the estimated total population is

$$\widehat{N} = a + b + c + (\frac{bc}{a+1}), \quad (3)$$

The estimated number of missed cases is

$$\hat{x} = bc/(a + 1), \quad (4)$$

and the 95% confidence intervals (95% CIs) of the estimated total population size $\widehat{N}$ is

$$95\%\ \text{CI} = \widehat{N} \pm 1.96\sqrt{\frac{(a+b+1)*(a+c+1)*b*c}{(a+1)^2*(a+2)}}. \quad (5)$$

3.1.2 MLRM Estimator

The MLRM estimator for estimating completeness in CR studies was first proposed by Sanathanan (Sanathanan, 1972) and later extended by Alho (1990). Define

indicator variables Y_{i1} and Y_{i2}, and m_i, for $i = 1, 2, \ldots, N$, where N is the unknown population size as defined in Table 3.1. Denote

$$Y_{ij} = \begin{cases} 1, & \text{if the } ith \text{ individual is captured in data source } j \text{ only}, j = 1, 2; \\ 0, & \text{otherwise;} \end{cases}$$

$$m_i = \begin{cases} 1, & \text{if the } i\text{th individual is captured by both data sources;} \\ 0, & \text{otherwise.} \end{cases}$$

Let $n_{ij} = Y_{ij} + m_i$, $M_i = Y_{i1} + Y_{i2} + m_i$, and define for the *i*th individual the probability of being captured in the *j*th data source as $p_{ij} = E[n_{ij}]$. Then $p_{i12} = E[m_i]$, where $p_{i12} = p_{i1}p_{i2}$ is the probability of being captured by both sources. By assuming that the probabilities of being captured are conditionally independent for the *i*th individual, we can define the model.

Let $\mathbf{X}_{ij}^{\top}$ be the value of the covariates associated with the *i*th individual in the *j*th data source where $^{\top}$ is the transpose operator, $\mathbf{X}_{i1} = (1, X_{i11}, \cdots, X_{i1k})^{\top}$; $\mathbf{X}_{i2} = (1, X_{i21}, \cdots, X_{i2h})^{\top}$. Then

$$(Y_{i1}, Y_{i2}, m_i, 1 - M_i) \sim \text{Mult}(1; p_{i1}(1 - p_{i2}), (1 - p_{i1})p_{i2}, p_{i12}, p_{0i}), \qquad (6)$$

where $p_{i0} = 1 - (p_{i1} + p_{i2} + p_{i12})$ is the probability that the *i*th individual is not identified in either data source. Let $\boldsymbol{\beta}_j = (\beta_{j0}, \cdots, \beta_{jk})^{\top}$ be the vectors of parameters associated with the covariates, and

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \boldsymbol{\beta}_j \, \mathbf{X}_{ij}^{\top} . \qquad (7)$$

Then $\hat{p}_{ij}$ is predicted from equation 7 by

$$\hat{p}_{ij} = \frac{e^{\hat{\boldsymbol{\beta}}_j \mathbf{X}_{ij}^{\top}}}{1+e^{\hat{\boldsymbol{\beta}}_j \mathbf{X}_{ij}^{\top}}} . \qquad (8)$$

Once the MLRM is fitted to the data and the probability of being captured for each observed patient is calculated, the sum of the reciprocal of these probabilities gives an estimate of the total population size, $\widehat{N}$, that is,

$$\widehat{N} = \sum_{M_i=1}\left(\frac{1}{1-\hat{p}_{i0}}\right), i = 1, \cdots, N, \tag{9}$$

where $\hat{p}_{i0}$ is the probability of being missed by all sources, and can be estimated as

$$\hat{p}_{i0} = \frac{1}{(1+e^{\hat{\beta}_1\tilde{x}_{i1}})(1+e^{\hat{\beta}_2\tilde{x}_{i2}})}. \tag{10}$$

The MLRM estimator can account for observable population heterogeneity in the capture probabilities. In other words, the characteristics of the captured individuals are used to explain their probabilities of capture.

The asymptotic variance of $\widehat{N}$ can be derived from a proposed estimator (Sekar & Deming, 1949):

$$V_1 = \frac{(a+c)*(a+b)*b*c}{a^3}. \tag{11}$$

This estimator does not account for the variability in the cases that are not captured. Alho (1990) presented an approximation to the unconditional variance which can be thought of as a generalization of V_1introduced by Sekar & Deming (1949). The unconditional variance estimator derived below can allow us to present unconditional confidence intervals for N under population heterogeneity even though a conditional likelihood was used in the estimation of $\boldsymbol{\theta}$, where $\boldsymbol{\theta}$ is a conditional maximum likelihood estimator.

Let $\boldsymbol{\Psi}(\boldsymbol{\theta}) = (\Psi_1(\boldsymbol{\theta}), \cdots, \Psi_{2N}(\boldsymbol{\theta}))^{\top}$, and then

$$\mathbf{V}(\boldsymbol{\theta}) = -\mathbf{X}^{\top}\boldsymbol{\Psi}(\boldsymbol{\theta}). \tag{12}$$

The formula for the estimator V_2 of the conditional asymptotic variance is

$$V_2 = \boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}})^{\top}\mathbf{X}(\mathbf{X}^{\top}\widehat{\boldsymbol{W}}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\top}\boldsymbol{\Psi}(\widehat{\boldsymbol{\theta}}), \quad (13)$$

where $\widehat{\boldsymbol{W}} = \text{cov}(\mathbf{Y}|\mathbf{M})$, and we defined $\mathbf{M} = (M_1, \cdots, M_N)^{\top}$, $\mathbf{Y} = (n_{11}, \cdots, n_{1N}, n_{21}, \cdots, n_{2N})^{\top}$.

Estimating the probability of being missed by both data sources gives the estimator V_3, which can be written as

$$V_3 = \sum_{M_i=1} \frac{p_{i0}}{(1-p_{i0})^2}. \quad (14)$$

Combining the results, we can get the unconditional estimator of $\text{var}(\widehat{N})$ to be

$$V_0 = V_2 + V_3. \quad (15)$$

3.2 Simulation Study

A Monte Carlo simulation study was undertaken to evaluate the performance of CR methods for estimating population size. The CR model proposed by Chapman (1951) was compared to the MLRM proposed by Alho (1990).

3.2.1 Data Generation

The Bernoulli distribution is used to describe experiments with dichotomous outcome variables (Kroese et al., 2013). The data for the simulation study were generated from a multivariate Bernoulli distribution, which allows manipulation of the magnitude of dependence between the variables. Specifically the random-variate vector $\mathbf{Y}_i = (Y_{i1}, Y_{i2})^{\top}$ was generated from a multivariate Bernoulli distribution with parameters $E(\mathbf{Y}_i) = (p_{i1}, p_{i2})^{\top}$ and $Corr(\mathbf{Y}_i) = \rho_{jj'}(j, j' = 1, 2)$ using the algorithm proposed by Emrich and Marion (Emrich & Piedmonte, 1991). The correlation matrix of multivariate normal data is used to produce binary vectors having the desired correlation.

In our study, $\rho = \rho_{12}$ had values of 0.0, 0.05, 0.1, 0.3, and 0.5 to investigate the effects of independence and increasing amounts of dependence between the data sources. To introduce heterogeneity into the capture probabilities, we adopted the following disease model when generating the data:

$$p_{ij} = \frac{1}{1+e^{-(\boldsymbol{\beta}_j \mathbf{x}_{ij}^{\top})}}. \quad (16)$$

where $\mathbf{X}_{ij} = (1, X_{ij1}, X_{ij2}, X_{ij3})^{\top}$ are the covariate vectors for the *i*th individual and the *j*th dataset and $\boldsymbol{\beta}_j = (\beta_{j0}, \beta_{j1}, \beta_{j2}, \beta_{j3})^{\top}$ are the vectors of model parameters. In this simulation, we assumed that X_{ij1} was a continuous and normally distributed covariate with parameters μ_1 and σ_1^2 denoting the mean and variance, respectively, and X_{ij2} and X_{ij3} were binary variables with parameters B_2, B_3 and p_{B_2}, p_{B_3} denoting the number of disease cases and corresponding event probabilities, respectively. The covariates were independently generated with $X_{ij1} \sim N(0,1)$ and $X_{ij2}, X_{ij3} \sim Bern(1,0.5)$. The parameters of the population disease model were selected in order to investigate a range of capture probability values. Misspecification of the MLRM across the seven scenarios was evaluated by replacing $\mathbf{X}_{ij} = (1, X_{ij1}, X_{ij2}, X_{ij3})^{\top}$ by $\mathbf{X}'_{ij} = (1, (X'_{ij1} + X''_{ij1}))^{\top}$, where $X'_{ij1}, X''_{ij1} \sim N(0,1)$, and by assuming that only X'_{ij1} was observed in the misspecified model. Accordingly, $\boldsymbol{\beta}'_j = (\beta_{j0}, \beta_{j1})^{\top}$ is the vector of model parameters for the *j*th dataset.

Three disease population sizes were considered with prevalence of 1%, 5%, and 10% in a population of 10,000. The completeness of each data source for capturing observed disease cases was manipulated, to look at the effect of unbalanced and balanced combinations of capture probabilities between two data sources. Different combinations

of capture probabilities between the two data sources resulted in different combinations of the vector of model parameters. Table 3.2 summarizes the relationships between the capture probability values and the vector of parameters. All our simulations were for the case of positive correlation between the data sources. Since a negative correlation is also a possibility, we conducted additional simulations under Scenario 1, Scenario 3, and Scenario 7 with a correlation of -0.10.

Table 3.2 Combinations of capture probabilities, vectors of MLRM parameters, E[n_1], E[n_2], E[m], and E[M] of the Monte Carlo simulation study

Scenario	**Model Parameters**								**E[n_1][a]**	**E[n_2][b]**	**E[m][c]**	**E[M][d]**
	β_{10}	β_{11}	β_{12}	β_{13}	β_{20}	β_{21}	β_{22}	β_{23}				
Scenario 1 (0.9, 0.9)[e]	2.21	-0.04 (0.03)[f]	0.04	-0.07	2.17	0.01 (-0.01)	-0.04	0.03	0.11*N*	0.09 *N*	0.79 *N*	0.99 *N*
Scenario 2 (0.9, 0.7)	2.30	0.04	-0.12	-0.05	0.90	0.02	-0.001	-0.02	0.27*N*	0.08*N*	0.63*N*	0.97*N*
Scenario 3 (0.9, 0.5)	2.11	0.002 (0.03)	0.04	0.11	-0.003	0.02 (0.02)	0.01	0.002	0.45*N*	0.06*N*	0.43*N*	0.94*N*
Scenario 4 (0.9, 0.2)	-0.01	-0.003	0.02	0.02	-2.20	0.008	0.05	-0.03	0.48*N*	0.06*N*	0.05*N*	0.59*N*
Scenario 5 (0.5, 0.5)	0.05	-0.007	-0.09	-0.02	0.04	0.03	-0.08	-0.05	0.25*N*	0.25*N*	0.23*N*	0.73*N*
Scenario 6 (0.5, 0.3)	0.04	-0.04	-0.02	-0.07	-0.79	0.007	-0.04	-0.02	0.33*N*	0.16*N*	0.17*N*	0.66*N*
Scenario 7 (0.5, 0.1)	-0.01	-0.003	0.02	0.02	-2.20	0.008 (0.05)	0.05	-0.03	0.45*N*	0.05*N*	0.04*N*	0.54*N*

[a]E[n_1] represents expected disease population being captured in the first data source only; [b]E[n_2] represents expected disease population being captured in the second data source only; [c]E[m] represents expected disease population being captured in both data sources; [d]E[M] represents expected disease population being captured in at least one of the data source; [e] These numbers represent disease prevalence in data source 1 and source 2. (0.9, 0.9) means that data source 1 has a disease prevalence of 90% and data source 2 has a disease prevalence of 90%; [f] The values in the briskets under $\beta_{11}$and β_{21} represent the coefficient values for the population parameters in the misspecification simulations if different from the values specified.

3.2.2 Statistical Analysis of Simulated Data

Two population size estimation methods were applied to each set of generated data: (a) Method 1: Chapman estimator (e.g., Chapman (1951)), which assumes homogeneity of capture probability and conditional independence of data sources, (b) Method 2: MLRM estimator with $\mathbf{X}_{ij} = (1, X_{ij1}, X_{ij2}, X_{ij3})^{\top}$ as covariates. In the model misspecification simulations, two models were applied to each set of generated data: (a) MLRM estimator with $\mathbf{X}'_{ij} = (1, (X'_{ij1} + X''_{ij1}))^{\top}$ as covariates (b) Method 3: MLRM estimator with $\mathbf{X}''_{ij} = (1, X'_{ij1})^{\top}$ as the sole covariate.

3.2.3 Measures of Model Performance

The estimated population size was computed for each combination of simulation conditions. Measures of relative bias (RB), coverage probability (CP) of 95% confidence intervals, width of the 95% CIs (WCI), and root-mean-square-error (RMSE) (Brittain & Böhning, 2009) were computed for each combination of simulation conditions and each replication. We defined RB $= \frac{(\widehat{N}-N)}{N} \times 100$, where $\widehat{N}$ is the estimated population size and N is the true population size; positive values indicate population size is overestimated, while negative values indicate underestimation. CP is the percentage of simulation replications in which the 95% CI captures the true population size. WCI is the difference between the lower and upper bounds of the 95% CI. RMSE is defined as the square root of the mean square error of an estimator $\hat{\theta}$ with respect to an estimated parameter θ, which is $\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} = \sqrt{\mathrm{E}((\hat{\theta} - \theta)^2)}$.

The Newton–Raphson method was used to estimate the coefficients of the covariates in the MLRM estimator and the RMSE was calculated for each set of

estimated coefficients between correctly-specified and misspecified models. Let $f(\beta)$ be a well-behaved function, and let t be a root of the equation $f(\beta) = 0$ (Ypma, 1995). The Newton-Raphson algorithm uses an iterative process to estimate parameters. If $\hat{\beta}_n$ is the current estimate, then the next estimate $\hat{\beta}_{n+1}$ is given by $\hat{\beta}_{n+1} = \hat{\beta}_n - \frac{f(\hat{\beta}_n)}{f'(\hat{\beta}_n)}$.

3.2.4 Simulation Organization

A total of 210 combinations of conditions were simulated. For each combination of conditions 1,000 replications were performed. For each replication the following steps were undertaken: (a) a set of data was generated (i.e., sampled) from the population with known characteristics using Scenario 1 to 7 to generate the data, (b) the proposed Chapman and MLRM estimators were applied to the data to estimate population size, and (c) each measure of model performance was computed. RB, WCI, and RMSE values were averaged across the 1,000 replications.

To investigate the effect of model misspecification on the MLRM estimators, we conducted simulations for conditions in which population disease prevalence was 10% (estimates were stable in large disease prevalence).We only examined Scenarios 1, 3, and 7 for the model misspecification conditions because they represent the scenarios of homogeneous capture probabilities, minimal heterogeneous capture probabilities, and moderate heterogeneous capture probabilities, respectively. We also conducted simulations where the correlation value was -0.10 because this is the most commonly used correlation value among previous simulation studies (Tilling & Sterne, 1999).

The simulation study was conducted using SAS/IML (Interactive Matrix Language) software version 9.3 (SAS Institute Inc, 2004), SAS/IML Studio (SAS

Institute Inc, 2013), and the R Project for Statistical Computing (The R Project for Statistical Computing, 2014).

3.3 Numeric Example

3.3.1 Study Design and Data Sources

Both the Chapman and MLRM estimators were applied to an existing dataset from the province of Saskatchewan, Canada. This dataset was originally created to compare estimates of RA and SARDs which include Systemic Lupus Erythematosus (SLE), Sjögren's syndrome (SjS), Systemic Sclerosis (SSc), Polymyositis (PM) and Dermatomyositis (DM) across multiple Canadian provinces and territories over a ten-year period (1998-2007) (Bernatsky et al., 2011).

Saskatchewan has a population of approximately 1.0 million according to the 2011 Statistics Canada Census (Statistics Canada, 2012). Like all Canadian provinces, Saskatchewan has a system of universal health care. All hospital records and virtually all physician billing records are captured for residents who are eligible to receive health insurance benefits. Individuals not eligible to receive provincial benefits include inmates in federal prisons, members of the national police service, and veterans, who represent about 1% of the population. Registered Indians, who represent about 9% of the population, do not receive provincial prescription drug benefits.

The two data sources that were used in this study to ascertain RA cases are hospital separation abstracts and physician billing claims, both of which are available in all provinces and territories in Canada. Thus, the methods that are developed here can be generalized to AHD from other jurisdictions in Canada.

Hospital discharge abstracts are completed upon discharge from an acute care facility and contain information on diagnosis and procedure codes, admission and discharge dates, length of stay, and service type (inpatient, day surgery, and outpatient). In Saskatchewan, prior to April 1, 2001, diagnoses were based on the International Classification of Diseases, 9th Revision (ICD-9). For hospital separations with a discharge date up to March 31, 1999, up to 3 diagnosis fields could be reported. For hospital separations with a discharge date from April 1, 1999 to March 31, 2001, up to 16 diagnoses could be reported. As of April 1, 2001, hospital discharge abstracts were changed to include 25 diagnosis codes based on the International Classification of Diseases and Related Health Problems, 10th Revision, Canada (ICD-10-CA).

Physicians who are paid on a fee-for-service basis submit billing claims to the provincial health ministry for payment purposes. A single diagnosis is recorded on each claim using three-digit International Classification of Diseases, 9th Revision (ICD-9) codes. Physicians can also submit claims for administrative purposes only (i.e., as a record of services provided) on alternate payment plans which are known as "shadow billing" (Manitoba Centre for Health Policy, 2008).

The population registration file was also used in this study; it captures dates of health insurance coverage, as well as information about demographic characteristics and location of residence. This data source was used to define covariates for the CR models. All data sources can be anonymously linked via a unique personal health number.

Cases of RA were identified using ICD-10-CA codes M05 and M06 and ICD-9 code 714. Individuals less than 19 years of age were excluded, to maintain a focus on the adult population. Data were available from January 1, 1998 to December 31, 2007 for

case ascertainment. Each case was assigned to a fiscal year by index date, which is the date of the first physician billing clam or hospital separation record with an RA diagnosis during the study period. The literature suggests that at least five years of AHD are required to obtain accurate incidence and prevalence estimates for other forms of arthritis (Bernatsky et al., 2009; Ng, Bernatsky, & Rahme, 2013; Ward, 2013) ; we have elected to use all data from the entire ten years of study data for RA to calculate the period prevalence.

A variable was created to identify the ascertainment source for each subject in the AHD as follows: diagnosis in hospital discharge abstracts only, diagnosis in physician billing claims only, and diagnoses in both data sources. Demographic variables available in the dataset that were used to describe the cohort and that may be associated with heterogeneity of capture probability include sex, age at time of first healthcare contact with RA diagnosis, and region of residence (e.g., urban area such as Saskatoon and Regina census metropolitan areas, rural area such as Lloydminster, Moose Jaw, and Prince Albert). Other diagnoses and measures of co-morbidity are not available in the provided data set.

3.3.2 Data Analysis

Both the Chapman and MLRM estimators were used to estimate the size of the RA population size in Saskatchewan. Model fit was assessed for the MLRM containing different sets of covariates (e.g., main effects only versus main and interaction effects) by using penalized measures of the log of the likelihood function, including the Akaike Information Criterion (Akaike, 1974), and the Bayesian-Schwarz Information Criterion (Schwarz, 1978). As well, the likelihood ratio test (LRT) which asymptotically follows a

χ^2 distribution was used to test the difference in fit for competing models with different sets of covariates.

3.4 Ethical Considerations

This study is part of a larger study that has already received ethical approval from the University of Saskatchewan, which is compliant with the Tri-Council Policy statement on Ethical Conduct for Research Involving Humans (see Appendix B). The University of Manitoba Health Research Ethics Board approved the request on continuing to use the same ethical approval for the thesis (see Appendix B).

CHAPTER 4 RESULTS

This chapter starts with the results from the simulation study to compare performance of the two CR methods using RB, CP, WCI, and RMSE. Next, the effects of model misspecification on the performance of the MLRM are described. This chapter concludes with a comparison of the two CR methods using data from the numeric example.

4.1 Monte Carlo Simulation Results

4.1.1 Comparison of the Chapman and MLRM Estimators

Table 4.1 presents results for RB in the estimated population size. When the assumptions of independence and homogeneity were satisfied (Scenario 1, correlation = 0), both the Chapman estimator $\widehat{N}$ and the MLRM estimator $\widehat{N}'$ produced RB estimates that were close to zero across all disease prevalence cases. When the data sources were independent, both of the estimators produced smaller RB values than when the data sources were dependent. However, one exception was that the estimated RB values from the MLRM estimator $\widehat{N}'$ became unstable and unpredictable when the data sources were independent and disease prevalence was low (1%).

Under all of the scenarios, the estimators $\widehat{N}$ and $\widehat{N}'$ were negatively biased when correlation existed between the two data sources. The estimates became more biased as the amount of correlation increased. For example, under Scenario 1 (disease prevalence = 10%) which was the homogeneous capture-probability case, the estimated RB values were close to 0 for both estimators. However, when the correlation increased to 0.5, the estimated RB values became increasingly negatively biased to -5.34%. Overall, the mean estimated RB values across all the scenarios were 0.00% and 0.69% respectively for the

Chapman and MLRM estimator when there was no correlation (disease prevalence = 5%). When the correlation increased to 0.3, the mean estimated RB values were -23.88% and -23.84% respectively for $\widehat{N}$ (ranging from -47.24% to -3.23%) and $\widehat{N}'$ (ranging from -47.18% to -3.23%).

As the data departed from the assumption of homogeneity of capture probability (Scenarios 4 and 7), the estimated population sizes became more biased for both estimators. The mean estimated RB values averaged across all the correlations in Scenario 1 and disease prevalence = 5% were the same (-1.23%) for the Chapman and MLRM estimator when the capture probabilities were homogeneous. The mean estimated RB values were -20.75% and -19.72% respectively for the two estimators $\widehat{N}$ and $\widehat{N}'$ in Scenario 4 (the capture probabilities were extreme heterogeneous) when the data were averaged across all correlation conditions.

When the assumptions of homogeneity of capture probability and independence of data sources were both violated, the estimated population sizes were more biased for both estimators. For example, from Scenario 1 to Scenario 4 (prevalence = 10%, correlation = 0.1) the estimated RB values were negatively biased from approximately -1.13% to -23.00% for both estimators. Note that for Scenarios 2, 3, 4, and 7, the estimated RB of population size could not be predicted when correlation was 0.5.

Disease prevalence also impacted the estimates when there was no correlation between the data sources. For example, the estimated RB values became unstable under Scenario 4 and Scenario 7 for the MLRM estimator when there was no correlation and low disease prevalence (1%). However, when the correlation value was minimal (e.g., 0.05, 0.1) and the capture probabilities were extremely heterogeneous (Scenario 7), the

MLRM estimator produced smaller RB values compared to the Chapman estimator when the disease prevalence was small (1%). Higher disease prevalence produced estimated RB values which were closer to zero for both of the estimators $\widehat{N}$ and $\widehat{N}'$. Overall, the estimated RB values were similar for the two estimators under different disease prevalence conditions when correlation existed between the data sources.

Table 4.2 contains results for the CP values. The CP values for both of the estimators were close to the nominal level of coverage (95%) when there was no correlation between the data sources under some scenarios (Scenario 1, disease prevalence = 5%; Scenario 3, disease prevalence = 10%). The estimated CP values were closest to the nominal level of coverage for all conditions with no correlation compared to conditions when correlation existed between the data sources. When the assumption of homogeneity of capture probability was satisfied, both estimators produced almost the same CP values (Scenario 1).

As the correlation increased, the CP values decreased dramatically for both of the estimators. Most of the estimated CP values were 0 when the correlation was 0.3. For example, the estimated CP was 92% under Scenario 1 when the data sources were independent (disease prevalence = 1%). And the estimated CP dropped to 2% when the correlation increased to 0.5 (Scenario 1, disease prevalence = 1%). Overall, the mean estimated CPs across all the scenarios were 93.29% and 94.24% respectively for the Chapman estimator (ranging from 89.90% to 95.20%, median = 93.30%) and the MLRM estimator (ranging from 93.00% to 95.30%, median = 94.30%) when there was no correlation (disease prevalence = 5%). At the same time, as the correlation increased to 0.3, the mean estimated CP values were 0.04% for both estimators.

When the assumption of homogeneity of capture probability was violated, the MLRM estimator $\widehat{N}'$, produced better CP estimates than the Chapman estimator, $\widehat{N}$ (Scenario 4 and Scenario 7). For example, under Scenario 4, the estimated CP was 91.90% for the MLRM estimator compared to 83.90% for the Chapman estimator (disease prevalence = 1%, correlation = 0). Overall, the mean estimated CP values across all the correlations in Scenario 1 (disease prevalence = 5%) were 51.30% and 51.40% for the Chapman estimator and the MLRM estimator respectively when the capture probabilities were homogeneous. At the same time, the mean estimated CPs were 43.68% and 46.65% respectively for the two estimators $\widehat{N}$ and $\widehat{N}'$ in Scenario 4 when the capture probabilities were extremely heterogeneous.

Disease prevalence seemed to impact on the estimated CP as correlation increased. As disease prevalence and correlation increased, the estimated CP values were further away from the nominal level of coverage (95%). For example, under Scenario 1 (correlation = 0.1), the estimated CP was almost 70.00% for both estimators when disease prevalence was 1%; this dropped to 37.00% and 15.00% respectively when disease prevalence was 5% and 10%. A disease prevalence of 1% produced larger CP values than a higher disease prevalence when there was correlation between data sources for both estimators. As disease prevalence increased when both of the dependence and homogeneity assumptions were violated, the estimated CP values were close to 0. Under Scenario 2, 3, 4, and 7 where the capture probabilities were heterogeneous, the CP values could not be estimated when the correlation between data sources was 0.5.

Table 4.3 contains the results for WCI values. The estimated CIs were wider when there was no correlation than when there was correlation between the data sources;

this result was found for both estimators. For example, under Scenario 1 (disease prevalence = 1%), the estimated CIs were 4.25 and 4.33 for each estimator respectively when the data sources were independent; the estimate CIs were only approximately 2.00 when the correlation was 0.5. Overall, the mean estimated WCIs across all the scenarios were 113.55 and 119.94 for the Chapman (ranging from 9.89 to 260.86) and MLRM estimators (ranging from 9.91 to 281.35, median = 89.58), respectively when there was no correlation (disease prevalence = 5%). As the correlation increased to 0.3, the mean estimated WCIs were 24.30 and 24.82, respectively, for the two estimators $\widehat{N}$ (ranging from 6.51 to 49.70, median = 27.24) and $\widehat{N}'$ (ranging from 6.53 to 50.38, median = 28.44).

The estimated CIs were wider for both of the estimators when the assumption of homogeneity was violated (Scenario 4 and Scenario 7) than when it was not violated. Overall, the mean estimated WCIs across all the correlations in Scenario 1 (disease prevalence = 5%) were 8.60 respectively for the Chapman estimator and the MLRM estimator when the capture probabilities were homogeneous. At the same time, the mean estimated WCIs were 154.76 and 165.14 respectively for the two estimators $\widehat{N}$ and $\widehat{N}'$ in Scenario 4 when the capture probabilities were extremely heterogeneous. Moreover, the estimated CIs were slightly wider for the MLRM estimator compared to the Chapman estimator when the capture probabilities were heterogeneous. For example, under Scenario 4 (disease prevalence = 1%, correlation = 0), the estimated CIs were 273.34 for the MLRM estimator and 104.66 for the Chapman estimator.

As disease prevalence increased, the confidence intervals became wider for all scenarios. For example, the mean estimated WCIs across all the correlations in Scenario

1 (disease prevalence = 1%) were 36.24 and 41.26, respectively, for the Chapman estimator and MLRM estimator. However, the mean estimated WCIs increased to 116.42 and 117.51 respectively for $\widehat{N}$ and $\widehat{N}'$ under the same scenario when the disease prevalence increased to 10%. The WCI values could not be predicted when the correlation was 0.5 for Scenario 2, 3, 4, and 7.

Although the estimated RB, CP, and WCI were similar for the two estimators $\widehat{N}$ and $\widehat{N}'$, the Chapman estimator $\widehat{N}$ resulted in increased difference between the estimated population and the true disease population for many conditions, as measured by RMSE (Table 4.4). Under all scenarios, the RMSE of the MLRM estimator $\widehat{N}'$ stayed stable as correlation and prevalence of disease increased. For example, under Scenario 1 (disease prevalence = 1%), the estimated average RMSE was approximately 10.00 across all the correlated cases for the MLRM estimator $\widehat{N}'$. However, the RMSE of the Chapman estimator $\widehat{N}$ increased significantly as the correlation increased. For example, under Scenario 1 (i.e., prevalence = 1%), the estimated RMSE was 1.30 when the two data sources were independent, and it increased to 33.10 as the correlation increased to 0.5 for the Chapman estimator $\widehat{N}$.

The smallest RMSE was observed for Scenario 1 with disease prevalence of 1% when the data sources were independent. Overall, the mean estimated RMSEs across all the scenarios were 1655.49 and 42.51 respectively for the Chapman estimator (ranging from 6.60 to 5256.30, median = 497.40) and the MLRM estimator (ranging from 22.20 to 80.00, median = 32.00) when there was no correlation (disease prevalence = 5%). At the same time, as the correlation increased to 0.3 where all the scenarios can converge, the mean estimated RMSEs were 21948.43 and 123.15 respectively for the two estimators $\widehat{N}$

(ranging from 281.90 to 55988.30, median = 13766.40) and $\widehat{N}'$ (ranging from 27.20 to 236.40, median = 118.00).

When the assumption of homogeneity was violated, the estimated RMSE became larger, especially for the Chapman estimator $\widehat{N}$. For example, the estimated RMSE was 10.10 under Scenario 1 (disease prevalence = 1%, correlation = 0.1) and it increased to 34.50 under Scenario 4 for the MLRM estimator $\widehat{N}'$. It increased from 3.30 to 900.60 for the Chapman estimator $\widehat{N}$ under the same scenario. As disease prevalence increased, the estimated RMSE for both estimators increased. The mean estimated RMSEs across all the correlations in Scenario 1 (disease prevalence = 5%) were 86.68 and 23.30, respectively, for the Chapman estimator and the MLRM estimator when the capture probabilities were homogeneous. At the same time, the mean estimated RMSEs were 20479.25 and 128.56 respectively for the two estimators $\widehat{N}$ and $\widehat{N}'$ in Scenario 4 when the capture probabilities were extremely heterogeneous.

Table 4.1 Relative bias (%) of estimated population size for Chapman estimator $\widehat{N}$ and MLRM estimator $\widehat{N}'$

Prevalence	Correlation	Scenario 1[1]		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6		Scenario 7	
		$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$
1.0%	**0.00**	0.03	0.04	-0.01	0.04	-0.02	0.13	-1.44	25.18	-0.44	0.89	0.69	4.09	1.34	27.92
	0.05	-0.56	-0.55	-0.98	-0.92	-1.66	-1.55	-12.36	4.63	-4.90	-3.80	-7.22	-4.64	-13.34	-0.79
	0.10	-1.09	-1.08	-2.20	-2.16	-3.20	-3.10	-22.86	-14.26	-9.35	-8.45	-13.12	-11.11	-22.72	-14.01
	0.30	-3.27	-3.26	-6.10	-6.08	-9.24	-9.22	-47.39	-45.83	-23.32	-22.92	-31.30	-30.61	-47.26	-45.57
	0.50	-5.26	-5.26	NA[2]	NA	NA	NA	NA	NA	-33.43	-33.27	-42.94	-42.73	NA	NA
5.0%	**0.00**	0.00	0.00	0.00	0.01	-0.01	0.01	-0.06	1.94	0.02	0.23	-0.22	0.27	0.30	2.36
	0.05	-0.57	-0.57	-1.06	-1.05	-1.67	-1.65	-12.91	-11.65	-4.78	-4.60	-6.64	-6.25	-13.24	-11.99
	0.10	-1.12	-1.12	-2.08	-2.08	-3.22	-3.20	-22.84	-22.03	-9.44	-9.30	-13.00	-12.71	-22.86	-22.05
	0.30	-3.23	-3.23	-6.00	-5.99	-9.15	-9.15	-47.17	-47.12	-23.21	-23.15	-31.13	-31.03	-47.24	-47.18
	0.50	-5.35	-5.35	NA	NA	NA	NA	NA	NA	-33.49	-33.47	-43.06	-43.04	NA	NA
10.0%	**0.00**	0.00	0.00	-0.02	-0.02	0.01	0.02	0.19	1.14	-0.11	-0.01	-0.14	0.09	-0.06	0.88
	0.05	-0.56	-0.56	-1.06	-1.06	-1.67	-1.66	-12.88	-12.28	-4.75	-4.66	-7.05	-6.87	-12.61	-12.01
	0.10	-1.13	-1.13	-2.08	-2.07	-3.21	-3.21	-22.97	-22.59	-9.21	-9.13	-13.09	-12.95	-22.81	-22.44
	0.30	-3.27	-3.27	-5.98	-5.98	-9.10	-9.10	-47.16	-47.13	-23.29	-23.26	-31.11	-31.06	-47.28	-47.25
	0.50	-5.34	-5.34	NA	NA	NA	NA	NA	NA	-33.65	-33.64	-42.84	-42.83	NA	NA

[1]For description of scenarios, see Table 3.2; [2] NA indicates an inestimable combination of conditions due to lack of model convergence.

Table 4.2 Coverage probability (%) of estimated population size for Chapman estimator $\widehat{N}$ and MLRM estimator $\widehat{N}'$

Prevalence	Correlation	Scenario 1[1]		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6		Scenario 7	
		$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$
1.0%	**0.00**	91.80	91.90	91.00	92.10	91.10	92.00	83.90	91.90	93.20	94.50	91.00	93.50	83.90	93.10
	0.05	82.50	83.60	83.20	84.60	82.10	82.90	70.10	82.60	79.90	84.60	79.00	84.10	68.90	83.10
	0.10	70.00	70.40	67.70	68.60	64.50	67.30	48.60	65.50	67.10	72.50	62.20	71.00	49.70	67.70
	0.30	19.60	19.60	8.20	8.60	1.30	1.80	0.30	3.70	5.00	7.10	4.20	6.30	0.20	3.40
	0.50	2.20	2.30	NA[2]	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	NA	NA
5.0%	**0.00**	95.20	95.30	94.20	94.30	92.90	93.10	92.90	95.00	94.60	95.10	93.30	93.90	89.90	93.00
	0.05	72.60	72.70	72.50	73.00	71.40	72.10	61.80	66.80	72.90	74.40	71.40	74.70	59.80	65.50
	0.10	37.10	37.30	32.60	33.10	27.80	28.40	20.00	24.80	26.70	28.40	28.00	30.60	20.50	23.90
	0.30	0.30	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.50	0.00	0.00	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	NA	NA
10.0%	**0.00**	94.10	94.00	95.30	95.20	95.00	95.10	92.60	93.70	94.00	93.90	94.80	94.70	93.30	94.80
	0.05	60.00	60.20	59.80	60.10	55.00	55.70	49.40	53.20	55.90	57.20	55.40	57.40	50.60	54.80
	0.10	15.00	15.00	9.70	9.90	7.00	7.20	4.70	5.90	6.80	7.40	8.30	9.10	5.00	5.60
	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.50	0.00	0.00	NA	NA	NA	NA	NA	NA	0.00	0.00	0.00	0.00	NA	NA

[1]For description of scenarios, see Table 3.2; [2] NA indicates an inestimable combination of conditions due to lack of model convergence.

Table 4.3 Width of 95% confidence intervals of estimated population size for Chapman estimator $\widehat{N}$ and MLRM estimator $\widehat{N}'$

Prevalence	Correlation	Scenario 1[1]		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6		Scenario 7	
		$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$
1.0%	**0.00**	4.25	4.33	8.11	8.31	12.72	13.22	104.66	273.34	38.51	41.65	58.96	68.43	111.70	265.72
	0.05	4.04	4.11	7.39	7.59	11.30	11.72	81.59	174.67	34.37	37.02	48.68	55.74	77.80	137.68
	0.10	3.83	3.91	6.82	6.99	9.72	10.12	59.60	99.10	30.04	32.20	42.03	47.57	60.07	100.42
	0.30	2.78	2.82	4.21	4.32	2.28	2.49	7.93	24.58	18.39	19.46	21.50	23.61	7.83	19.00
	0.50	1.91	1.94	NA[2]	NA	NA	NA	NA	NA	10.51	11.03	10.05	10.94	NA	NA
5.0%	**0.00**	9.89	9.91	18.55	18.63	29.12	29.30	256.86	276.73	88.43	89.58	131.15	134.10	260.86	281.35
	0.05	9.22	9.25	17.08	17.15	25.99	26.15	191.34	203.75	78.25	79.20	112.78	115.09	189.84	202.10
	0.10	8.71	8.74	15.72	15.78	23.02	23.15	143.34	151.39	68.29	69.06	95.23	97.00	143.67	151.71
	0.30	6.51	6.53	9.99	10.03	7.57	7.65	27.48	28.68	41.62	42.00	49.70	50.38	27.24	28.44
	0.50	4.52	4.54	NA	NA	NA	NA	NA	NA	24.04	24.21	23.18	23.42	NA	NA
10.0%	**0.00**	14.00	14.02	26.24	26.29	41.46	41.58	368.98	382.29	124.66	125.43	186.38	188.40	366.04	379.19
	0.05	13.16	13.18	24.21	24.26	36.95	37.06	272.45	280.74	110.57	111.22	157.34	158.88	273.96	282.26
	0.10	12.37	12.39	22.23	22.28	32.56	32.65	203.84	209.17	97.68	98.22	134.77	136.00	204.43	209.76
	0.30	9.29	9.31	14.23	14.26	11.08	11.11	40.82	41.52	58.93	59.19	70.79	71.26	41.01	41.70
	0.50	6.39	6.39	NA	NA	NA	NA	NA	NA	34.01	34.13	32.83	33.00	NA	NA

[1]For description of scenarios, see Table 3.2; [2]NA indicates an inestimable combination of conditions due to lack of model convergence.

Table 4.4 Root-mean-squared-error of estimated population size for Chapman estimator $\widehat{N}$ and MLRM estimator $\widehat{N}'$

Prevalence	Correlation	Scenario 1[1]		Scenario 2		Scenario 3		Scenario 4		Scenario 5		Scenario 6		Scenario 7	
		$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$	$\widehat{N}$	$\widehat{N}'$
1.0%	**0.00**	1.3	10.0	4.7	9.9	10.9	10.8	1133.2	117.4	100.9	14.4	262.3	20.7	1268.5	103.3
	0.05	2.0	10.2	6.0	10.0	13.6	10.1	803.9	69.9	114.1	14.3	236.3	18.4	647.6	38.1
	0.10	3.3	10.1	10.4	10.2	20.6	10.9	900.6	34.5	159.1	15.3	318.3	19.6	843.2	31.8
	0.30	14.3	10.1	44.8	11.5	96.3	13.1	2312.1	51.35	599.3	25.2	1046.4	32.4	2265.6	46.3
	0.50	33.1	11.0	NA[2]	NA	NA	NA	NA	NA	1161.9	34.5	1896.0	43.5	NA	NA
5.0%	**0.00**	6.6	22.2	24.3	23.1	61.7	23.2	4544.9	76.3	497.4	32.0	1197.2	40.8	5256.3	80.0
	0.05	16.6	21.8	54.6	22.2	121.7	23.9	6984.9	83.4	996.7	37.3	1987.9	48.1	6999.1	83.0
	0.10	41.6	22.0	134.0	23.9	313.1	27.1	14504.1	118.3	2560.7	53.7	4894.8	71.6	14660.2	118.7
	0.30	281.9	27.2	942.0	37.0	2146.3	50.3	55883.1	236.23	13766.4	118.0	24631.0	156.9	55988.3	236.4
	0.50	748.5	34.0	NA	NA	NA	NA	NA	NA	28079.0	167.9	46612.6	215.9	NA	NA
10.0%	**0.00**	13.6	30.1	43.8	31.6	111.5	31.6	10147.8	107.5	1081.2	43.2	2227.8	55.1	9034.7	102.4
	0.05	48.7	29.4	158.0	32.9	386.2	34.2	21723.6	144.7	3117.6	62.4	6778.2	84.6	20930.7	142.7
	0.10	149.8	33.5	479.8	36.3	1134.5	44.6	55529.7	233.2	9151.4	98.8	18553.2	137.0	54921.1	232.0
	0.30	1107.7	44.3	3647.1	66.7	8376.7	95.6	222108.5	471.08	54786.8	234.8	97454.0	312.3	224426.6	473.6
	0.50	2911.7	60.8	NA	NA	NA	NA	NA	NA	113834.8	337.8	183489.9	428.3	NA	NA

[1]For description of scenarios, see Table 3.2; [2] NA indicates an inestimable combination of conditions due to lack of model convergence.

Table 4.5 contains the results from the simulations when the correlation value was -0.1. In general, the presence of negative source dependence resulted in overestimation of the population size for both of the CR estimators. Under Scenario 1, in which the assumption of homogeneity of capture probabilities was satisfied, both estimators $\widehat{N}$ and $\widehat{N}'$ produced RB values that were slightly larger than 0. For this same scenario, CP values were 1.80% and 1.70%, respectively for the two estimators. The WCI was similar between the two estimators. The RMSE was larger in the conventional estimator $\widehat{N}$ compared to the MLRM estimator $\widehat{N}'$. As the data departed more from the assumption of homogeneity of capture probability the RB values were more biased and always overestimated the total population size (Scenario 7). As well, the 95% CIs became wider, and the RMSE became larger. However, the CP values increased to the nominal level as heterogeneity existed in capture probabilities (i.e., Scenario 1 vs. Scenario 7).

Table 4.5 Performance of Chapman estimator $\widehat{N}$ and MLRM estimator $\widehat{N}'$ when disease prevalence = 10% and correlation = -0.10

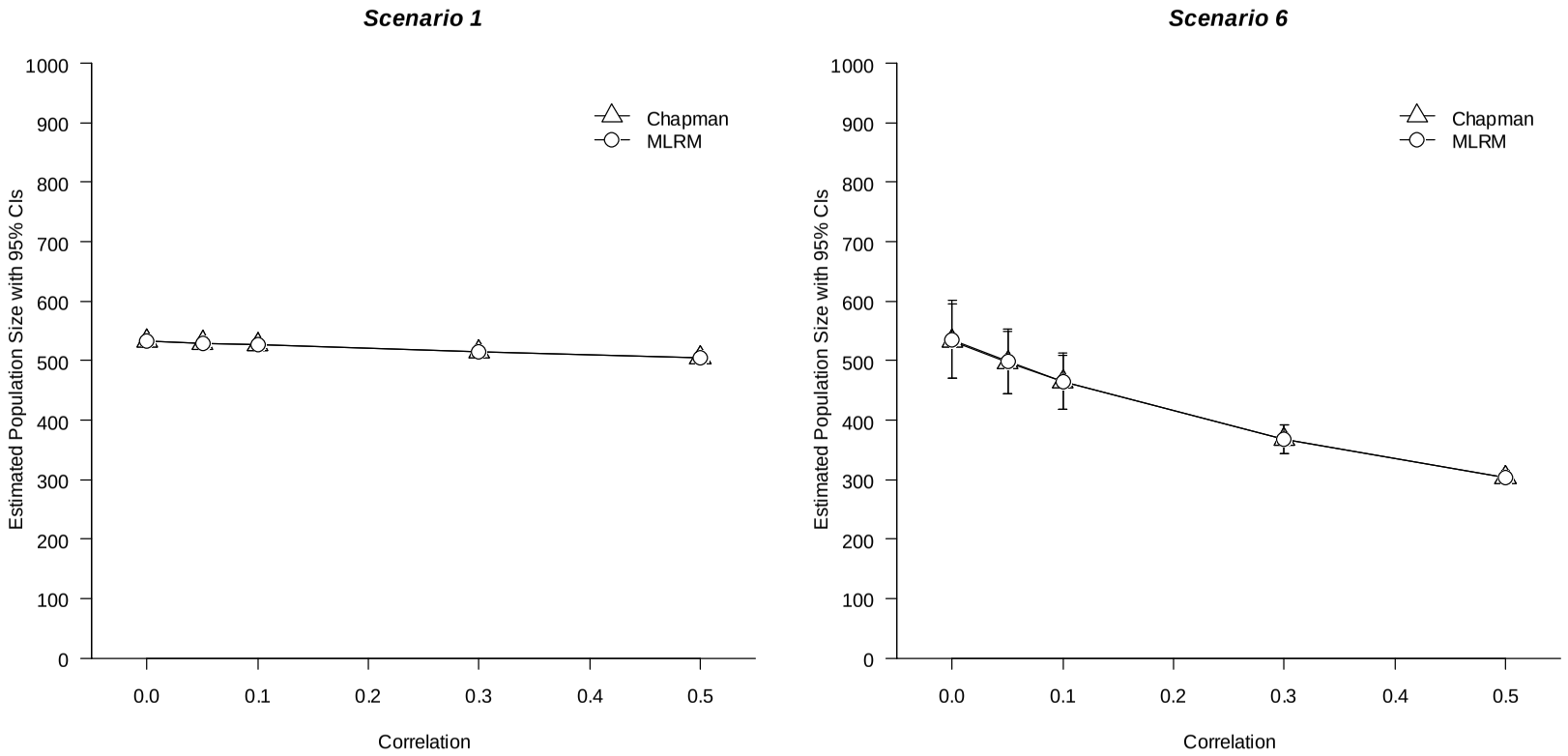
	Method	Scenario 1[1]	Scenario 3	Scenario 7
RB[2] (%)	$\widehat{N}$	1.15	3.50	41.99
	$\widehat{N}'$	1.15	3.51	44.53
CP[3] (%)	$\widehat{N}$	1.80	17.00	27.80
	$\widehat{N}'$	1.70	16.70	25.80
WCI[4]	$\widehat{N}$	15.65	50.81	715.96
	$\widehat{N}'$	15.68	50.97	754.90
RMSE[5]	$\widehat{N}$	136.46	1371.64	212177.57
	$\widehat{N}'$	33.15	49.13	490.16

Note: [1]For description of scenarios, see Table 3.2; [2]RB is the relative bias; [3]CP is the percentage of replications in which the 95% CI captures the true population size; [4]WCI is the difference between the lower and upper bounds of the 95% CI; [5]RMSE is the root-mean-square-error.

Figure 4.1 shows the average estimated population size with 95% CIs across disease prevalence conditions for the Chapman and MLRM estimators. We compared Scenario 1 and Scenario 6 because Scenario 1 represents homogeneity in capture probabilities and Scenario 6 represents moderate heterogeneity in capture probabilities. Furthermore, both scenarios converged when the correlation was 0.5. The average 95% CIs did not differ between the two estimators across each of the scenario.

When the two data sources were independent, both of the estimators produced wider 95% CIs. When the assumption of independence of data sources was violated, both of the estimators produced narrower 95% CIs. For Scenario 1, when the assumption of homogeneity was met, all the cases with various correlations had the narrowest 95% CIs of estimated population size across all disease prevalence conditions. For Scenario 6, when the assumption of homogeneity was violated, all the estimated population sizes had much wider 95% CIs compared to those in Scenario 1.

Figure 4.1 Estimated population size and 95% confidence intervals (95% CIs) for Scenarios 1 and 6 as correlation increased



Note: Chapman refers to the Chapman estimator $\widehat{N}$; MLRM refers to the MLRM estimator $\widehat{N}'$; for description of scenarios, see Table 3.2; lower error bar = 5^{th} percentile; upper error bar = 95^{th} percentile of the distribution.

4.1.2 Effects of Model Misspecification for the MLRM

The results found in Table 4.6 show the effect of model misspecification on performance of the MLRM for estimating population size. The estimated population size did not differ for the correctly-specified model and misspecified model in terms of RB, CP, WCI, and RMSE across the scenarios. However, model misspecification resulted in biased parameter estimates.

Under Scenario 1 the population parameters were $\beta_{11} = 0.01$ and $\beta_{21} = -0.03$. When the model was correctly specified, we obtained average estimates of $\hat{\beta}_{11} = 0.01$ and $\hat{\beta}_{21} = -0.03$ with RMSE values of 0.09 and 0.05, respectively, when there was no correlation between the data sources. Under the misspecified model for Scenario 1, the average estimates were $\hat{\beta}_{11} = 0.001$ and $\hat{\beta}_{21} = -0.03$ with RMSE values of 0.13 and 0.07, respectively, when there was no correlation. When the correlation increased, both of the parameter estimates became slightly biased for the correctly specified model and the misspecified model for both $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$. However, the parameter estimates were similar for the independence case and the dependence case. The RMSE values were similar for the two estimated coefficient values when the correlation increased.

Under Scenario 7 the population parameters were $\beta_{11} = -0.03$ and $\beta_{12} = 0.03$. When the model was correctly specified, the coefficient estimates were equivalent to their parameter values and the RMSE values were 0.11 and 0.08, respectively, when there was no correlation between the data sources. Under the misspecified model, the estimated coefficients were similar to their corresponding parameters but the RMSE values were 0.16 and 0.12, respectively, when there was no correlation. As the correlation increased, both of the parameter estimates became more biased for both the correctly specified and

the misspecified model for $\hat{\beta}_{11}$. However, $\hat{\beta}_{21}$ did not seem to be affected as the correlation increased for both the correctly-specified model and the misspecified model because the estimated coefficient value remained stable and the RMSE didn't change, either.

Across all scenarios, when the assumption of independence was violated, the estimated parameters became increasingly biased for both the correctly-specified model and the misspecified model. For example, the RMSE for $\hat{\beta}_{11}$ increased as correlation increased. However, there was no difference between the RMSE for the correctly-specified model and the misspecified model as correlation increased for each single scenario for $\hat{\beta}_{21}$.When the assumption of homogeneity was violated, the estimated parameters became more biased for both the correctly-specified model and the misspecified model. For example, the RMSE for $\hat{\beta}_{11}$was higher for Scenario 7 (the heterogeneous case) compared to Scenario 1 (the homogeneous case). There was no difference between the RMSE for the correctly-specified model and the misspecified model as correlation increased for each scenario. However, the RMSE for $\hat{\beta}_{21}$ increased slightly for Scenario 7 compared to Scenario 1.

Table 4.6 Performance of the MLRM estimator without misspecification ($\widehat{N}'$) and with misspecification ($\widehat{N}''$) for Scenarios 1, 3, and 7 when disease prevalence is 10%

	Method	Scenario 1[1]			Scenario 3			Scenario 7		
		Correlation								
		0.0	**0.1**	**0.3**	**0.0**	**0.1**	**0.3**	**0.0**	**0.1**	**0.3**
RB[2] (%)	$\widehat{N}'$	-0.04	-2.08	-6.00	0.01	-3.18	-8.96	1.04	-23.11	-46.73
	$\widehat{N}''$	-0.04	-2.08	-6.00	0.01	-3.18	-8.96	1.03	-23.11	-46.73
CP[3] (%)	$\widehat{N}'$	93.50	9.80	0.00	94.00	7.60	0.00	94.30	6.00	0.00
	$\widehat{N}''$	93.60	9.90	0.00	94.00	7.70	0.00	94.30	6.00	0.00
WCI[4]	$\widehat{N}'$	26.25	22.32	14.02	40.62	31.88	10.83	392.77	210.75	42.40
	$\widehat{N}''$	26.25	22.32	14.02	40.62	31.88	10.83	392.75	210.75	42.40
RMSE[5]	$\widehat{N}'$	32.09	35.28	66.77	32.18	44.17	94.13	106.82	238.78	467.75
	$\widehat{N}''$	32.09	35.28	66.77	32.17	44.17	94.13	106.79	238.78	467.75
RMSE_$\widehat{\beta}_{11}$[6]	$\widehat{N}'$	0.09	0.10	0.15	0.11	0.12	0.34	0.11	0.14	0.35
	$\widehat{N}''$	0.13	0.14	0.21	0.15	0.17	0.50	0.16	0.20	0.51
RMSE_$\widehat{\beta}_{12}$[7]	$\widehat{N}'$	0.05	0.05	0.05	0.05	0.05	0.05	0.08	0.09	0.08
	$\widehat{N}''$	0.07	0.07	0.08	0.07	0.07	0.07	0.12	0.12	0.11

Note: [1]For description of scenarios, see Table 2; [2]RB is the relative bias; [3]CP is the percentage of replications in which the 95% CI captures the true population size; [4]WCI is the difference between the lower and upper bounds of the 95% CI; [5]RMSE is the root-mean-square deviation; [6]RMSE_$\hat{\beta}_{11}$ is the root-mean-square deviation of the estimated parameter β_{11}; [7] RMSE_$\hat{\beta}_{12}$ is the root-mean-square deviation of the estimated parameter β_{12}.

4.2 Numeric Example

Table 4.7 contains some descriptive results for the RA dataset. There were a total of 19,290 RA cases (19+ years) identified in the Saskatchewan population during the 10-year period from 1998 to 2007. Overall, 83.3% of cases were identified from diagnosis codes in physician billing claims only, 4.1% of cases were identified from diagnosis codes in hospital separation abstracts only, and 12.6% of cases were identified from diagnosis codes in both data sources. Females (67.0%) were more likely to be found in all of the data sources compared to males (33.0%). People 35 years and older (90.0%) were more likely to be captured in all of the data sources compared to people who were younger (9.96%). As well, cases that lived in urban areas (63.9%) were more likely to be captured in all of the data sources compared to those who live in rural areas (36.1%).

Male patients were more likely to be ascertained in physician data (34.1%) compared to hospital data (29.3%) and both data sources (27.0%). Female were more likely to be ascertained in hospital data (70.7%) compared to physician data (65.9%). People who were aged 19 to 54 were more likely to be ascertained in physician data (47.1%) than hospital data (18.1%). People who were 55 to 74 years old were more likely to be ascertained in both data sources (45.7%) than physician data only (34.9%) or in hospital data only (32.8%). People who were 75 years and older were more likely to be ascertained in hospital data only (49.1%) compared to physician data only (18.1%). In terms of the residence, RA cases were ascertained almost equally in three data sources not only for those who lived in the urban area but also for those who lived in the rural area.

Table 4.7 Frequency (%) of RA cases captured in diagnosis codes from AHD by demographic variables and data source

Variable		Data Source		
		Physician data only (83.3)	Hospital data only (4.1)	Both sources (12.6)
Sex	Male	5484 (34.1)	235 (29.3)	655 (27.0)
	Female	10584 (65.9)	563 (70.7)	1769 (73.0)
Age Group	19-34	1719 (10.7)	39 (4.9)	163 (6.7)
	35-54	5847 (36.4)	105 (13.2)	577 (23.8)
	55-74	5601 (34.9)	262 (32.8)	1107 (45.7)
	>=75	2901 (18.1)	392 (49.1)	577 (23.8)
Residence	Urban	10145 (63.1)	518 (64.9)	1665 (68.7)
	Rural	5923 (36.9)	280 (35.1)	759 (31.3)
Total		16068	798	2424

Table 4.8 shows the distribution of RA cases across the study years by data source. Most RA cases were ascertained from physician billing claims in each study year. Since we only identified RA cases starting from the index year of 1998, there were more RA cases ascertained in that year for both data sources. Beginning in 2004, hospital discharge abstracts only and both data sources captured almost the same number of RA cases.

Table 4.8 Frequency (%) of RA cases captured in diagnosis codes from AHD across index year and data source

Index year	Physician data only	Hospital data only	Both sources	Total
1998	2911 (65.4)	161 (3.5)	1389 (31.1)	4461
1999	1678 (79.1)	128 (6.4)	320 (15.5)	2126
2000	1652 (85.1)	85 (4.0)	204 (10.9)	1941
2001	1464 (86.4)	89 (5.7)	136 (7.9)	1689
2002	1437 (90.0)	69 (4.0)	103 (6.0)	1609
2003	1467 (90.5)	61 (3.6)	89 (5.9)	1617
2004	1446 (91.5)	67 (4.9)	64 (3.6)	1577
2005	1354 (92.0)	50 (4.0)	51 (4.0)	1455
2006	1343 (94.6)	47 (2.7)	43 (2.7)	1433
2007	1316 (95.8)	41 (2.8)	25 (1.4)	1382

Table 4.9 contains model fit statistics and parameter estimates for the MLRM estimator for different sets of covariates. The following models were fit to the data: a) Model 1: sex, age group, residence; 2) Model 2: sex, age group, residence, sex*age group; 3) Model 3: sex, age group, residence, age group*residence; 4) Model 4: sex, age group, residence, sex*residence. We only use the two-way interactions because three-way interactions could not always be fit to the data because of sparse cell sizes. The reference category for the dependent variable is being captured in the physician billing claims. The reference categories for the independent variables are male, less than 35 years old and rural residence.

Identification of the optimal model based on fit statistics suggests that only very small improvements in AIC are achieved when interaction effects are added to the model (Model 2). As well, there are few differences in model fit statistics amongst the models with different interaction terms. For the BIC, the main effects model (Model 1) has the best fit. The likelihood ratio test statistics were less than 0.01 in size when the main effects model was compared to models that contained interaction effects. The residence variable was found not to be significant at the 5% level for hospital discharge abstracts for all models.

The total number of RA cases identified from the data without using a CR method was 19290. By using the Chapman estimator, we estimated the population size to be 24577 (95% CI: 24123, 25031), an increase of 27.4%. In contrast, using the MLRM estimator for the main effects model, we estimated the population size to be 20118 (95% CI: 19664, 20572), an increase of 4.3%. To calculate period prevalence, we used the 2006 Statistics Canada Census data (>=19 years) as the denominator from the province of

Saskatchewan (Statistics Canada, 2012). The estimated RA prevalence was 2.76% when prevalence was based on the number of cases captured in hospital and physician claims only, 3.52% for the Chapman estimator, and 2.88% for the MLRM estimator.

Table 4.9 Model fit statistics and parameter estimates (standard errors) from the MLRM estimator with different sets of covariates

		Model			
		1[1]	**2**[2]	**3**[3]	**4**[4]
Model fit statistics	**AIC**	20300.5	20295.9	20300.0	20299.6
	BIC	20394.9	20437.5	20441.6	20409.7
	-2LogL	20276.5	20259.9	20264.0	20271.6
Parameter estimates (Standard errors)	$\widehat{\boldsymbol{\beta}}_{\mathbf{intercept1}}$[5]	2.29 (0.08)	2.77 (0.18)	2.23 (0.10)	-2.23 (0.08)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{intercept2}}$	-3.99 (0.17)	-3.97 (0.31)	-4.08 (0.23)	3.92 (0.18)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{sex1}}$	-0.31 (0.04)	-0.93 (0.20)	-0.31 (0.04)	0.23 (0.05)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{sex2}}$	0.18 (0.08)	0.15 (0.36)	0.18 (0.08)	-0.08 (0.10)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{age1}}$	-0.58 (0.08)	-1.09 (0.18)	-5.2 (0.10)	0.58 (0.08)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{age2}}$	0.79 (0.17)	0.77 (0.31)	0.89 (0.23)	-0.79 (0.17)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{residence1}}$	0.19 (0.04)	0.19 (0.04)	0.33 (0.16)	-0.37 (0.08)
	$\widehat{\boldsymbol{\beta}}_{\mathbf{residence2}}$	-0.03 (0.08)	-0.03 (0.08)	0.19 (0.32)	0.23 (0.14)
Pr < ChiSq	$\widehat{\boldsymbol{\beta}}_{\mathbf{intercept1}}$	<0.01	<0.01	<0.01	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{intercept2}}$	<0.01	<0.01	<0.01	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{sex1}}$	<0.01	<0.01	<0.01	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{sex2}}$	0.02	0.67	0.02	0.39
	$\widehat{\boldsymbol{\beta}}_{\mathbf{age1}}$	<0.01	<0.01	<0.01	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{age2}}$	<0.01	0.01	<0.01	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{residence1}}$	<0.01	<0.01	0.03	<0.01
	$\widehat{\boldsymbol{\beta}}_{\mathbf{residence2}}$	0.69	0.69	0.57	0.11

[1] Model 1: sex, age group, residence; [2] Model 2: sex, age group, residence, sex*age group;[3] Model 3 sex, age group, residence, age group*residence; [4] Model 4: sex, age group, residence, sex*residence.[5] $\beta_{\text{intercept1}}$ represents the constant for physician billing claims.

CHAPTER 5 DISCUSSION AND CONCLUSIONS

5.1 Summary

In this study, we investigated the performance of the Chapman estimator and the MLRM estimator for population size in the two-source CR problem under data-analytic conditions characterized by dependence between data sources and heterogeneity of capture probabilities, which are likely to arise when AHD are used to estimate chronic disease prevalence in AHD. We conducted both a Monte Carlo simulation study and a numeric example. In addition, the effect of model misspecification was examined for the MLRM estimator. We chose to focus on the Chapman estimator and the MLRM estimator because these two estimators are the most commonly used CR methods from the literature (Alho, 1990; Tilling & Sterne, 1999).

Under scenarios in which the two data sources were not correlated, the Chapman estimator slightly underestimated the population size and the MLRM estimator slightly overestimated the population size, but the amount of bias in these estimators was small. Under scenarios in which the data sources were correlated, both of the CR methods underestimated the population size and became more biased as the amount of correlation increased. However, the estimates were almost the same for both of the CR methods when correlation existed.

CP values were closest to the nominal level of coverage (i.e., 95%) when there was no correlation and became increasingly smaller as correlation increased. Overall, both of the CR methods produced almost the same estimates in terms of RB, CP, and WCI for each combination of the investigated simulation conditions. However, the RMSE increased dramatically for the Chapman estimator when correlation increased

from 0.0 to 0.5 compared to the MLRM estimator, which produced similar estimates of RMSE across all combinations of simulation conditions.

When the capture probabilities were heterogeneous, both of the estimators $\widehat{N}$ and $\widehat{N}'$ produced larger RB, wider confidence intervals, and larger RMSE compared to the homogeneous case. For example, Scenario 1 (the homogeneous case) resulted in the smallest RB, larger CP, closest WCI, and smallest RMSE among all scenarios for both estimators. However, the MLRM estimator produced better CP values than the Chapman estimator when capture probabilities were heterogeneous.

Model misspecification did not result in differences in terms of the performance of the MLRM estimates. However, as expected, the parameter estimates were biased in the misspecified model compared to the correctly-specified model. This is partly consistent with the work of Alho (1990)'s. Alho (1990) proposed that the misspecified model would perform worse than the original model in the main simulation study. However, he did not directly compare estimates from the misspecified model and the correctly-specified model.

Our simulations were primarily conducted with positive correlation between the captures. Since a negative correlation is also a possibility, we also investigated selected condition in which the correlation was negative. Our results showed that both of the estimators overestimated the population size when the correlation was negative. And this overestimation could be extreme when the capture probabilities were heterogeneous. At the same time, the estimated CIs of the population size became wider, and RMSE value became larger when capture probabilities were heterogeneous. However, one exception

was that CP values were closer to the nominal coverage when capture probabilities were heterogeneous and data sources were negatively dependent.

Computing times were similar and efficient for each scenario (i.e., approximately three hours). The MLRM estimator $\widehat{N}'$ took a slightly longer time to produce compared to the Chapman estimator $\widehat{N}$. As the correlation increased, both of the estimators took a longer time to compute.

The estimated RA prevalence was slightly higher in the MLRM estimator than the crude estimate when we applied the CR estimators to the numeric example from Saskatchewan. The Chapman estimator was larger than the RA prevalence estimate than the MLRM estimator. There was no difference among models including various covariates in terms of the population size estimates under the MLRM estimator but the parameter estimates did differ at 5% level for model specifications. However, the RA prevalence based the CR methods in the numeric example may be biased according to the simulation study. Regardless of the unmeasurable correlation between the two data sources, we had 4.1% RA cases being ascertained in hospital data only and only 12.6% of cases were captured in both data sources. The amount of overlap between the data sources may have affected the estimates of RA prevalence.

Our simulation results show some similarities with results from previous research. Wittes (1972) proposed that Chapman's estimator was unbiased when the assumption of homogeneity of capture probabilities was satisfied. Alho (1990) proposed that the classical CR method (see equation 1) underestimated population size while the MLRM estimator overestimated population size when there was correlation between two data sources. When the assumption of homogeneity was violated, the estimates were more

biased for both of the estimators. We got similar estimates for RB, CP, and WCI for both the Chapman and the MLRM estimator. Alho (1990) proposed that the conditional estimate of is asymptotically equivalent to the MLE estimator (the Chapman estimator) which was also proposed by Sanathanan (1972). Tilling & Sterne (1999) also found that the estimated CIs narrowed as positive correlation increased. Our research is also consistent with the literature (Brenner, 1995) in terms of that CR methods in two dependent sources underestimate the population size if the two sources are positively dependent, and overestimate the population size if the two sources are negatively dependent. However, none of the above research considered the violations of independence of data sources and homogeneity of capture probabilities simultaneously, which make our study unique.

5.2 Strength and Limitations

Our study has a number of strengths. First of all, we explored various combinations of source dependence and homogeneous/heterogeneous capture probabilities. The method that we used to generate the binary correlated data is very important in illustrating the assumption of independence between data sources. We investigated correlation values ranging from -0.1 to 0.5, although not all the combinations of simulation conditions could be investigated when the correlation between data sources was 0.5. Compared to previous simulation studies in the literature, our study has the advantage of choosing not only a wider range but also larger values of correlation between data sources (Tilling & Sterne, 1999). Also, we introduced observable variability in capture probabilities via covariate effects which allowed us to examine the violations

of the assumption of homogeneity of capture probabilities. The capture probabilities that we used were ranging from homogeneous one to extreme heterogeneous one.

Secondly, we used two-source CR methods in our study because physician billing claims and hospital discharge abstracts are the most common AHD in Canadian provinces. Three or more sources can also be used in CR problems, so researchers might also consider CR methods for multiple AHD. Thirdly, we manipulated the number of disease cases (i.e., prevalence) in the simulation study to investigate the effect of sparsity in disease cases on the performance of CR methods. Fourthly, we considered misspecification of the MLRM to explore the covariate effects closely. We not only looked at the performance of CR methods including covariate effects but also looked at the model misspecification effects. Finally, we applied the CR methods in a numeric example to demonstrate its application in real-world data.

This study also has some limitations. Firstly, we used a conditional variance estimate, which Alho (1990) proposed, to estimate the confidence interval for the MLRM estimate of population size. The literature suggests that the conditional variance estimate will be similar to the variance estimate for Chapman's method (Alho, 1990; Sanathanan, 1972); we might also have used an empirical bootstrap technique for estimating the variance. The bootstrap estimator might produce narrower confidence intervals. It involves generating an empirical distribution for the estimated population size by randomly sampling with replacement from the original dataset, estimating $\widehat{N}$ in each random sample, and repeating this process multiple times. In general, at least 1000 bootstrap samples are recommended to attain good precision (Efron & Tibshirani, 1994). The empirical values from these bootstrap samples are rank-ordered from smallest to

largest and the 2.5th percentile and 97.5th percentiles of the distribution are used to approximate the lower and upper bounds of the 95% CI. However, Tilling & Sterne (1999) found that the coverage of the bootstrap confidence interval was consistently lower than the nominal coverage of 95%.

In addition, we only had a limited set of covariates available in our numeric example. This limited our analysis in terms of model specification when estimating population size. However, according to both our simulation study and numeric example, the results suggest that model misspecification does not have a large biasing effect on population size.

5.3 Conclusions and Future Work

In conclusion, we have compared the Chapman estimator and the MLRM estimator to estimate population size from two AHD sources. We introduced dependence of captures and heterogeneity of capture probabilities in a simulation study simultaneously. One of the key assumptions of CR methods is the independence of data sources, which is often violated in real-world data. As a result, researchers who wish to use CR methods for both of the estimators should be careful when negative source dependence is of concern because CR methods will overestimate the population sizes. Other than that, CR methods could be valuable to correct for underascertainment of cases even if positive source dependence exists.

Furthermore, researchers should have sufficient overlap (e.g., 50% of the cases are captured by both data sources) between the data sources when using CR methods for estimating population size in order to minimize heterogeneity of capture probabilities. Researchers should also consider linking multiple data sources such as AHD and survey

data, or different sources of AHD when using CR methods. Two data sources may result in a sparse number of cases in one data source (e.g., hospital discharge abstracts) for rare conditions such as RA. AHD such as prescription drug data can also be linked to physician billing claims for identifying rare conditions such as RA. However, validations studies of sensitivity and specificity in identifying diseases need to be conducted before applying CR methods. For example, prescription drug data may lack in specificity in identifying RA which may not be an ideal data source for estimating RA prevalence.

When we compared the two estimators, the MLRM estimator produced better CP values and smaller RMSE values than the Chapman estimator when the capture probabilities were heterogeneous. According to our results, missing covariates did not impact on the prevalence estimates in the MLRM estimator. Given these research findings, researchers who wish to estimate the size of chronic disease populations using two AHD should adopt the MLRM estimator instead of the Chapman estimator when covariate information is available in the data.

Although direct testing of the assumptions of CR models, especially the assumption of independence of data sources, is not possible, Brenner (1995) suggested that for the application of CR methods to epidemiologic monitoring of disease, information about the healthcare contact behavior of individuals with a specific design can often provide insights about the likely direction and magnitude of correlation. For example, more severe cases, residence within the registration area, and easy access to medical care can lead to positive dependence of between data sources. Negative dependence will exist when case ascertainment by different sources may be "mutually exclusive" (Brenner, 1995). This arises when different hospitals are used as separate

sources for case ascertainment under CR models and patients are more likely to be treated in one hospital than another. Brenner (1995) also suggested that negative dependence may exist between sources such as pathology and hematology laboratory results for ascertaining malignancies such as leukemia and lymphomas, because these sources are less common for advanced cancers with poorer prognosis.

Brenner (1995) suggested several strategies may help to minimize the degree of dependence between data sources such as the definition of sources, stratified analyses, and using three sources of case ascertainment. In terms of choosing the best sources like AHD to be included in the CR models, we found that we need to consider both the percentage of overlap of the data sources and potential correlation. For RA, using physician billing claims and hospital discharge abstracts may lead to a positive dependence for ascertaining RA because severe patients may be referred to a rheumatologist or be hospitalized. However, since we had so few cases identified in hospital discharge abstracts only, the percentage of overlapping cases was extremely low, which affects the accuracy of estimation. Other AHDs might be used for case ascertainment, such as prescription drug data.

Our future work should compare different methods for deriving confidence intervals for CR methods. For example, a non-parametric approach could be used to be compared with the conditional variance estimator that we adopted. At the same time, we should also consider including validation values in our simulation study to check the performance of CR methods in prevalence and incidence estimates. Because this would be more appropriate in the real-world data. Finally, we could conduct additional simulation studies to examine the performance of three-source CR methods.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on, 19*(6), 716-723. doi: 10.1109/TAC.1974.1100705

Alho, J. M. (1990). Logistic regression in capture-recapture models. *Biometrics, 46*(3), 623-635.

Barton, J. L., Trupin, L., Schillinger, D., Gansky, S. A., Tonner, C., Margaretten, M., . . . Yelin, E. (2011). Racial and ethnic disparities in disease activity and function among persons with rheumatoid arthritis from university-affiliated clinics. *Arthritis Care and Research, 63*(9), 1238-1246. doi: 10.1002/acr.20525

Bernatsky, S., Joseph, L., Pineau, C. A., Bélisle, P., Boivin, J. F., Banerjee, D., & Clarke, A. E. (2009). Estimating the prevalence of polymyositis and dermatomyositis from administrative data: Age, sex and regional differences. *Annals of the Rheumatic Diseases, 68*(7), 1192-1196. doi: 10.1136/ard.2008.093161

Bernatsky, S., Lix, L., Hanly, J. G., Hudson, M., Badley, E., Peschken, C., . . . Joseph, L. (2011). Surveillance of systemic autoimmune rheumatic diseases using administrative data. *Rheumatology International, 31*(4), 549-554. doi: 10.1007/s00296-010-1591-2

Bernillon, P., Lievre, L., Pillonel, J., Laporte, A., & Costagliola, D. (2000). Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993. *International Journal of Epidemiology, 29*(1), 168-174.

Bombardier, C., Hawker, G., & Mosher, D. (2011). The impact of arthritis in Canada: today and the next 30 years. Toronto: Arthritis Alliance of Canada.

Brenner, H. (1995). Use and limitations of the capture-recapture method in disease monitoring with two dependent sources. *Epidemiology, 6*(1), 42-48.

Brittain, S., & Böhning, D. (2009). Estimators in capture-recapture studies with two sources. *AStA Advances in Statistical Analysis, 93*(1), 23-47. doi: 10.1007/s10182-008-0085-y

Chapman, D.G. (1951). *Some properties of the hypergeometric distribution with applications to zoological sample censuses*: University of California Press.

Efron, B., & Tibshirani, R.J. (1994). *An Introduction to the Bootstrap*: Taylor & Francis.

Emrich, Lawrence J., & Piedmonte, Marion R. (1991). A Method for Generating High-Dimensional Multivariate Binary Variates. *The American Statistician, 45*(4), 302-304. doi: 10.1080/00031305.1991.10475828

Gabriel, S. E. (2001). The epidemiology of rheumatoid arthritis. *Rheumatic Disease Clinics of North America, 27*(2), 269-281.

Giarrizzo, M. L., Pezzotti, P., Silvestri, I., & Di Lallo, D. (2007). Estimating prevalence of diabetes mellitus in a Lazio province, Italy, by capture-recapture models. *Stima di prevalenza di diabete mellito in una provincia del lazio attraverso i modelli cattura e ricattura., 31*(6), 333-339.

Health Canada. (2012). *Canadian Community Health Survey*. Ottawa: Retrieved from http://www.hc-sc.gc.ca/fn-an/surveill/nutrition/commun/index-eng.php.

Hebert, P. L., Geiss, L. S., Tierney, E. F., Engelgau, M. M., Yawn, B. P., & McBean, A. M. (1999). Identifying persons with diabetes using medicare claims data. *American Journal of Medical Quality, 14*(6), 270-277.

Helmick, C. G., Felson, D. T., Lawrence, R. C., Gabriel, S., Hirsch, R., Kwoh, C. K., . . . Stone, J. H. (2008). Estimates of the prevalence of arthritis and other rheumatic conditions in the United States. Part I. *Arthritis and Rheumatism, 58*(1), 15-25. doi: 10.1002/art.23177

Hook, E. B., & Regal, R. R. (1995). Capture-recapture methods in epidemiology: Methods and limitations. *Epidemiologic Reviews, 17*(2), 243-264.

Kabasakal, Y., Kitapcioglu, G., Turk, T., Öder, G., Durusoy, R., Mete, N., . . . Akalin, T. (2006). The prevalence of Sjögren's syndrome in adult women. *Scandinavian Journal of Rheumatology, 35*(5), 379-383. doi: 10.1080/03009740600759704

Kriegsman, D. M. W., Penninx, B. W. J. H., Van Eijk, J. Th M., Boeke, A. J. P., & Deeg, D. J. H. (1996). Self-reports and general practitioner information on the presence of chronic diseases in community dwelling elderly. A study on the accuracy of patients' self-reports and on determinants of inaccuracy. *Journal of Clinical Epidemiology, 49*(12), 1407-1417. doi: 10.1016/S0895-4356(96)00274-0

Kroese, D.P., Taimre, T., & Botev, Z.I. (2013). *Handbook of Monte Carlo Methods*: Wiley.

Lincoln, F.C. (1930). *Calculating Waterfowl Abundance on the Basis of Banding Returns*: U.S. Department of Agriculture.

Lipscombe, L. L., & Hux, J. E. (2007). Trends in diabetes prevalence, incidence, and mortality in Ontario, Canada 1995-2005: a population-based study. *Lancet, 369*(9563), 750-756. doi: 10.1016/S0140-6736(07)60361-4

Lix, L. M., Yogendran, M. S., Leslie, W. D., Shaw, S. Y., Baumgartner, R., Bowman, C., . . . James, R. C. (2008). Using multiple data features improved the validity of

osteoporosis case ascertainment from administrative databases. *Journal of Clinical Epidemiology, 61*(12), 1250-1260. doi: 10.1016/j.jclinepi.2008.02.002

Lix, L. M., Yogendran, M. S., Shaw, S. Y., Burchill, C., Metge, C., & Bond, R. (2008). Population-based data sources for chronic disease surveillance. *Chronic Diseases in Canada, 29*(1), 31-38.

Lix, L., Yogendran, M., Burchill, C., Mettge, C., McKeen, N., Moore, D. , & Bond, R. (2006). Defining and Validating Chronic Diseases: An Administrative Data Approach. Winnipeg: Manitoba Centre for Health Policy.

Manitoba Centre for Health Policy. (2007). Term: Health Survey. from http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=102760

Manitoba Centre for Health Policy. (2008). Term: Shadow Billing. from http://mchp-appserv.cpe.umanitoba.ca/viewDefinition.php?definitionID=103569

Manuel, D. G., Rosella, L. C., & Stukel, T. A. (2010). Importance of accurately identifying disease in studies using electronic health records. *BMJ (Clinical research ed.), 341*. doi: 10.1136/bmj.c4226

McClish, D., & Penberthy, L. (2004). Using multivariate capture-recapture techniques and statewide hospital discharge data to assess the validity of a cancer registry for epidemiologic use. *Health Services and Outcomes Research Methodology, 5*(2), 141-152. doi: 10.1007/s10742-005-4305-6

Michaud, K., & Wolfe, F. (2007). Comorbidities in rheumatoid arthritis. *Best Practice and Research: Clinical Rheumatology, 21*(5), 885-906. doi: 10.1016/j.berh.2007.06.002

Ng, R., Bernatsky, S., & Rahme, E. (2013). Observation period effects on estimation of systemic lupus erythematosus incidence and prevalence in Quebec. *Journal of Rheumatology, 40*(8), 1334-1336. doi: 10.3899/jrheum.121215

Peragallo, M. S., Urbano, F., Lista, F., Sarnicola, G., & Vecchione, A. (2011). Evaluation of cancer surveillance completeness among the Italian army personnel, by capture-recapture methodology. *Cancer Epidemiology, 35*(2), 132-138. doi: 10.1016/j.canep.2010.06.016

Petersen, C. (1896). The yearly immigration of young plaice into the Limfjord from the German sea. *The Danish Biologicai Station 6*, 1-48.

Rasch, E. K., Hirsch, R., Paulose-Ram, R., & Hochberg, M. C. (2003). Prevalence of rheumatoid arthritis in persons 60 years of age and older in the United States: Effect of different methods of case classification. *Arthritis and Rheumatism, 48*(4), 917-926. doi: 10.1002/art.10897

Sanathanan, L. (1972). Estimating the Size of a Multinomial Population. *The Annals of Mathematical Statistics, 43*(1), 142-152.

SAS Institute Inc. (2004). *SAS/IML ® 9.1 User's Guide* Retrieved from http://support.sas.com/documentation/onlinedoc/91pdf/sasdoc_91/iml_ug_7306.pdf

SAS Institute Inc. (2013). *SAS/IML® Studio 12.3: User's Guide*

Schmidtmann, I. (2008). Estimating completeness in cancer registries - Comparing capture-recapture methods in a simulation study. *Biometrical Journal, 50*(6), 1077-1092. doi: 10.1002/bimj.200810483

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461-464. doi: 10.2307/2958889

Sekar, C. Chandra, & Deming, W. Edwards. (1949). On a Method of Estimating Birth and Death Rates and the Extent of Registration. *Journal of the American Statistical Association*, *44*(245), 101-115. doi: 10.2307/2280353

Shapira, Y., Agmon-Levin, N., & Shoenfeld, Y. (2010). Geoepidemiology of autoimmune rheumatic diseases. *Nature Reviews Rheumatology*, *6*(8), 468-476. doi: 10.1038/nrrheum.2010.86

Simard, J. F., & Mittleman, M. A. (2007). Prevalent rheumatoid arthritis and diabetes among NHANES III participants aged 60 and older. *Journal of Rheumatology*, *34*(3), 469-473.

Singh, J. A., Holmgren, A. R., & Noorbaloochi, S. (2004). Accuracy of veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis Care and Research*, *51*(6), 952-957. doi: 10.1002/art.20827

Spasoff, R. A. (1995). Health department administration of the Canadian Health Care Program. *Journal of Public Health Policy*, *16*(2), 141-151. doi: 10.2307/3342590

Statistics Canada. (2009). *Canadian Community Health Survey (CCHS) - Cycle 1.1*. Ottawa: Retrieved from http://www.statcan.gc.ca/concepts/health-sante/index-eng.htm.

Statistics Canada. (2012). *Saskatoon, Saskatchewan (Code 4711066) and Saskatchewan (Code 47) (table). Census Profile. 2011 Census.* Ottawa: Statistics Canada Retrieved from http://www12.statcan.gc.ca/census-recensement/2011/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CSD&Code1=4711066&Geo2=PR&C

ode2=47&Data=Count&SearchText=Saskatoon&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=4711066.

Statistics Canada. (2013). *Canadian Community Health Survey - Annual Component (CCHS)*. Ottawa: Retrieved from http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226.

Suissa, S., & Garbe, E. (2007). Primer: Administrative health databases in observational studies of drug effects - Advantages and disadvantages. *Nature Clinical Practice Rheumatology, 3*(12), 725-732. doi: 10.1038/ncprheum0652

Symmons, D., Turner, G., Webb, R., Asten, P., Barrett, E., Lunt, M., . . . Silman, A. (2002). The prevalence of rheumatoid arthritis in the United Kingdom: New estimates for a new century. *Rheumatology, 41*(7), 793-800.

The R Project for Statistical Computing. (2014). The R Project for Statistical Computing. from http://www.r-project.org/

Tilling, K., & Sterne, J. A. C. (1999). Capture-recapture models including covariate effects. *American Journal of Epidemiology, 149*(4), 392-400.

Tilling, K., Sterne, J. A. C., & Wolfe, C. D. A. (2001). Estimation of the incidence of stroke using a capture-recapture model including covariates. *International Journal of Epidemiology, 30*(6), 1351-1359.

Toronto Western Research Institute. (2010). Prevalence of Arthritis and Rheumatic Diseases around the World: A Growing Burden and Implications for Health Care Needs. Toronto: Toronto Western Research Institute.

van hest, N. A. H., Smit, F., Baars, H. M., de vries, G., de haas, P. E. W., Westenend, P. J., . . . Richardus, J. H. (2007). Completeness of notification of tuberculosis in

The Netherlands: How reliable is record-linkage and capture - Recapture analysis? *Epidemiology and Infection, 135*(6), 1021-1029. doi: 10.1017/S0950268806007540

Van Walraven, C., Austin, P. C., Manuel, D., Knoll, G., Jennings, A., & Forster, A. J. (2010). The usefulness of administrative databases for identifying disease cohorts is increased with a multivariate model. *Journal of Clinical Epidemiology, 63*(12), 1332-1341. doi: 10.1016/j.jclinepi.2010.01.016

Virnig, B. A., & McBean, M. (2001) Administrative data for public health surveillance and planning. *Vol. 22* (pp. 213-230).

Waltz, M., Kriegel, W., & Van't Pad Bosch, P. (1998). The social environment and health in rheumatoid arthritis: Marital quality predicts individual variability in pain severity. *Arthritis Care and Research, 11*(5), 356-374.

Ward, M. M. (2013). Estimating disease prevalence and incidence using administrative data: Some assembly required. *Journal of Rheumatology, 40*(8), 1241-1243. doi: 10.3899/jrheum.130675

Wennberg, J., & Gittelsohn, A. (1973). Small area variations in health care delivery. A population based health information system can guide planning and regulatory decision making. *Science, 182*(4117), 1102-1108.

Widdifield, J., Bernatsky, S., Paterson, J. M., Tu, K., Ng, R., Thorne, J. C., . . . Bombardier, C. (2013). Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: A validation study using the medical records of rheumatologists. *Arthritis Care and Research, 65*(10), 1582-1591. doi: 10.1002/acr.22031

Wittes, Janet T. (1972). 331. Note: On the Bias and Estimated Variance of Chapman's Two-Sample Capture-Recapture Population Estimate. *Biometrics, 28*(2), 592-597. doi: 10.2307/2556173

Wunsch, H., Harrison, D. A., & Rowan, K. (2005). Health services research in critical care using administrative data. *Journal of Critical Care, 20*(3), 264-269. doi: 10.1016/j.jcrc.2005.08.002

Yang, N. P., Deng, C. Y., Chou, Y. J., Chen, P. Q., Lin, C. H., Chou, P., & Chang, H. J. (2006). Estimated prevalence of osteoporosis from a Nationwide Health Insurance database in Taiwan. *Health Policy*, *75*(3), 329-337. doi: 10.1016/j.healthpol.2005.04.009

Yip, P. S. F., Bruno, G., Tajima, N., Seber, G. A. F., Buckland, S. T., Cormack, R. M., . . . McCarty, D. J. (1995). Capture-recapture and multiple-record systems estimation I: History and theoretical development. *American Journal of Epidemiology, 142*(10), 1047-1058.

Young, T. K. (2005). *Population Health: Concepts and Methods 2nd Edition*. New York: Oxford University Press.

Ypma, Tjalling J. (1995). Historical development of the Newton-Raphson method. *SIAM Review*, *37*(4), 531-551.

APPENDIX: R PROGRAMS

Two sample programs are provided for the Chapman estimator and the MLRM estimator in applications of simulated data.

/*----------*/

Completeness of Rheumatoid Arthritis Prevalence Estimates from Administrative Health Databases: Comparison of Capture-Recapture Models

Copyright(c) 2014 by Yao Nie

/*----------*/

Programmer: Yao Nie

Date: 2014.04.17

Scenario: 1

If you have further request, you can send email to: niey@myumanitoba.ca

/*----------*/

1 Chapman Estimator

```
Proc IML ;
**Specify simulation parameters**
nsim = 1000 ; *Number of simulations*
correlation = 0 ; *Correlation value*
alpha = 0.05 ;
**Specify counters for population estimators**
N_True=j(nsim,1) ; *True disease population size*
CH_N=j(nsim,1) ; *Estimated disease population size*
V_CH1=j(nsim,1)  ;
```

```
V_CH2=j(nsim,1) ;
V_CH=j(nsim,1)  ; *Variance of estimated disease population size*
Bound_CH=j(nsim,2)  ; *Lower and Upper bounds of estimated disease population size*
Coverage_CH=j(nsim,1) ; *Coverage probability of estimated disease population size*
WCI_CH=j(nsim,1) ; *Width of 95% confidence intervals of estimated disease
population size*
**Main body of Monte Carlo simulation** ;
do j =1 to nsim ;
submit / R;
 #Generate correlated binary data
 install.packages('mvtBinaryEP')
 library(mvtBinaryEP)
 #Total number of observations in the dataset
 n = 10000
 #Specify the coefficient values
 beta01 = 2.21
 beta11 = -0.035
 beta21 = 0.043
 beta31= -0.069
 beta02 =  2.17
 beta12 =  0.01
 beta22 = -0.043
 beta32= 0.0297
```

```
beta03 = -25

beta13 = 0.5

beta23 = 0

beta33= 0

#Specify counters for population estimators

y = NULL

x1 = NULL

x2 = NULL

d = NULL

p = NULL

for (i in 1:n)

{

 #Specify the disease prevalence = 1%

di = rbinom(1, 1, 0.1)

 #Create true diseased population status

if (di == 1){

#Generate two sets of a continuous covariate and two binary covariates

xi11 = rnorm(1, mean = 0, sd = 1)

xi21 = rbinom(1,1,0.5)

xi31 = rbinom(1,1,0.5)

xi12 = rnorm(1, mean = 0, sd = 1)

xi22 = rbinom(1,1,0.5)

xi32 = rbinom(1,1,0.5)
```

```
 pi1 =
exp(beta01+beta11*xi11+beta21*xi21+beta31*xi31)/(1+exp(beta01+beta11*xi11+beta2
1*xi21+beta31*xi31))
 pi2 =
exp(beta02+beta12*xi12+beta22*xi22+beta32*xi32)/(1+exp(beta02+beta12*xi12+beta2
2*xi22+beta32*xi32))
  #Specify correlations between the data sources
 mu = c(pi1, pi2)
 R = c(1, 0, 0, 1)
 R = matrix(R, ncol=2)
 ep0 = ep(mu=mu, R=R, nRep=1, seed=NULL)
 y0 = ep0$y
 }
#Create the rest of the true population
 else {
 xi11 = rnorm(1, mean = 0, sd = 1)
 xi21 = rbinom(1,1,0.5)
 xi31 = rbinom(1,1,0.5)
 xi12 = rnorm(1, mean = 0, sd = 1)
 xi22 = rbinom(1,1,0.5)
 xi32 = rbinom(1,1,0.5)
 pi1 = 0
```

```
 pi2 =
exp(beta03+beta13*xi11+beta23*xi21+beta33*xi31)/(1+exp(beta03+beta13*xi11+beta2
3*xi21+beta33*xi31))
 yi1 = 0
 yi2 = rbinom(1,1,pi2)
 y0 = c(yi1, yi2)
 pi = c(pi1,pi2)
 }
 xi1 = c(1, xi11, xi21, xi31)
 xi2 = c(1, xi12, xi22, xi32)
 x1 = rbind(x1, xi1)
 x2 = rbind(x2, xi2)
 y = rbind(y,y0)
 d = rbind(d, di)
 p = rbind(p,pi)
 }
 #Counts the total number of diseased individual from the true population in data source
1 only
  n10=0
 for (i in 1:n){
 if (y[i,1]==1 & y[i,2]==0)
  {n10=n10+1
  }
```

```
 }
 #Counts the total number of diseased individual from the true population in data source
2 only
 n01=0
 for (i in 1:n){
 if (y[i,1]==0 & y[i,2]==1)
  {n01=n01+1
  }
 }
 #Counts the total number of diseased individual from the true population in both data
sources
 n11=0
 for (i in 1:n){
 if (y[i,1]==1 & y[i,2]==1)
  {n11=n11+1
  }
 }
endsubmit;
**Import matrix from R**
run ImportMatrixFromR(n10,'n10');
run ImportMatrixFromR(n01,'n01');
run ImportMatrixFromR(n11,'n11');
run ImportMatrixFromR(d,'d');
```

```
**Calculate the true diseased population size**
N_True[j,]=sum(d) ;
**Chapman estimator for the estimated population size**
CH_N[j,]=(n10+n11+1)*(n01+n11+1)/(n11+1)-1 ;
**Chapman estimator for measures of performance for the estimated population size**
V_CH1[j,]=(n10+n11+1)*(n01+n11+1)*n10*n01 ;
V_CH2[j,]=(n11+1)*(n11+1)*(n11+2) ;
V_CH[j,]=V_CH1[j,]/V_CH2[j,] ;
Bound_CH[j,1]=CH_N[j,]+(-Probit(1-alpha/2))*sqrt(V_CH[j,]) ;
Bound_CH[j,2]=CH_N[j,]+(Probit(1-alpha/2))*sqrt(V_CH[j,]) ;
Coverage_CH[j,]=(Bound_CH[j,1]<=N_True[j,])*(Bound_CH[j,2]>=N_True[j,]) ;
WCI_CH[j,]=Bound_CH[j,2]-Bound_CH[j,1] ;
end ;
**Calculate the average of true diseased population size**
Mean_N=N_True[:,];
**Calculate the average of true diseased population size**
Mean_CH_N=CH_N[:,];
**Calculate the average measures of performance for Chapman estimator**
RB_CH=(Mean_CH_N-Mean_N)/Mean_N ;
CP_CH=Coverage_CH[:,];
Mean_WCI_CH=WCI_CH[:,];
Mean_Bound_CH= Bound_CH[:,];
MSE_CH_N=sum((CH_N-N_True)##2)/nsim;
```

```
**Combine results**

Result_CH=nsim||correlation||Mean_N||Mean_CH_N||RB_CH||CP_CH||Mean_WCI_CH||

Mean_Bound_CH||MSE_CH_N;

**Print out results**

Print Result_CH ;
```

2 MLRM Estimator

```
Proc IML ;
**Specify simulation parameters**
nsim = 1000 ; *Number of simulations*
correlation = 0 ; *Correlation value*
alpha = 0.05 ;
**Specify counters for population estimators**
N_True=j(nsim,1) ; *True disease population size*
ML3_Theta=j(nsim,8) ;
ML3_N=j(nsim,1) ; *Estimated disease population size*
ML3_V=j(nsim,1) ; *Variance of estimated disease population size*
Bound_ML3=j(nsim,2) ; *Lower and Upper bounds of estimated disease population
size*
Coverage_ML3=j(nsim,1) ; *Coverage probability of estimated disease population size*
WCI_ML3=j(nsim,1) ; *Width of 95% confidence intervals of estimated disease
population size*
ML3_beta01_est=j(nsim,1) ; *Estimated parameters from MLEM estimator*
ML3_beta11_est=j(nsim,1) ;
ML3_beta21_est=j(nsim,1) ;
ML3_beta31_est=j(nsim,1) ;
ML3_beta02_est=j(nsim,1) ;
ML3_beta12_est=j(nsim,1) ;
ML3_beta22_est=j(nsim,1) ;
```

```
ML3_beta32_est=j(nsim,1) ;
**Main body of Monte Carlo simulation**
do j =1 to nsim ;
submit / R;
 #Generate correlated binary data
 install.packages('mvtBinaryEP')
 library(mvtBinaryEP)
 #Total number of observations in the dataset
 n = 10000
 #Specify the coefficient values
 beta01 = 2.21
 beta11 = -0.035
 beta21 = 0.043
 beta31= -0.069
 beta02 =  2.17
 beta12 =  0.01
 beta22 = -0.043
 beta32= 0.0297
 beta03 = -25
 beta13 = 0.5
 beta23 = 0
 beta33= 0
 y = NULL
```

```
x1 = NULL
x2 = NULL
d = NULL
p = NULL
for (i in 1:n)
{
#Specify the disease prevalence = 1%
di = rbinom(1, 1, 0.1)
#Create true diseased population status
if (di == 1){
#Generate two sets of a continuous covariate and two binary covariates
xi11 = rnorm(1, mean = 0, sd = 1)
xi21 = rbinom(1,1,0.5)
xi31 = rbinom(1,1,0.5)
xi12 = rnorm(1, mean = 0, sd = 1)
xi22 = rbinom(1,1,0.5)
xi32 = rbinom(1,1,0.5)
pi1 =
exp(beta01+beta11*xi11+beta21*xi21+beta31*xi31)/(1+exp(beta01+beta11*xi11+beta2
1*xi21+beta31*xi31))
pi2 =
exp(beta02+beta12*xi12+beta22*xi22+beta32*xi32)/(1+exp(beta02+beta12*xi12+beta2
2*xi22+beta32*xi32))
```

```
#Specify correlations between the data sources
mu = c(pi1, pi2)
R = c(1, 0, 0, 1)
R = matrix(R, ncol=2)
ep0 = ep(mu=mu, R=R, nRep=1, seed=NULL)
y0 = ep0$y
}
#Create the rest of the true population
else {
xi11 = rnorm(1, mean = 0, sd = 1)
xi21 = rbinom(1,1,0.5)
xi31 = rbinom(1,1,0.5)
xi12 = rnorm(1, mean = 0, sd = 1)
xi22 = rbinom(1,1,0.5)
xi32 = rbinom(1,1,0.5)
pi1 = 0
pi2 =
exp(beta03+beta13*xi11+beta23*xi21+beta33*xi31)/(1+exp(beta03+beta13*xi11+beta2
3*xi21+beta33*xi31))
yi1 = 0
yi2 = rbinom(1,1,pi2)
y0 = c(yi1, yi2)
```

```
 pi = c(pi1,pi2)
 }
 xi1 = c(1, xi11, xi21, xi31)
 xi2 = c(1, xi12, xi22, xi32)
 x1 = rbind(x1, xi1)
 x2 = rbind(x2, xi2)
 y = rbind(y,y0)
 d = rbind(d, di)
 p = rbind(p,pi)
 }
endsubmit;
**Import matrix from R**
run ImportMatrixFromR(beta01,'beta01');
run ImportMatrixFromR(beta11,'beta11');
run ImportMatrixFromR(beta21,'beta21');
run ImportMatrixFromR(beta31,'beta31');
run ImportMatrixFromR(beta02,'beta02');
run ImportMatrixFromR(beta12,'beta12');
run ImportMatrixFromR(beta22,'beta22');
run ImportMatrixFromR(beta32,'beta32');
run ImportMatrixFromR(x1,'x1');
run ImportMatrixFromR(x2,'x2');
run ImportMatrixFromR(y,'y');
```

```
run ImportMatrixFromR(d,'d');

run ImportMatrixFromR(p,'p');

z = y||d||x1||x2||p;

**Calculate the true diseased population size**

N_True[j,]=sum(d) ;

create a from z;

append from z;

delete all where (col1=0 & col2=0);

purge;

use a ;

read all into obs;

close a;

**Reshape matrix from a different dimensions**

ones=shape(1,nrow(obs),1);

Zeros=shape(0,nrow(obs),1);

Y1=obs[,{1}];Y2=obs[,{2}];

X1=Ones||obs[,{5}]||obs[,{6}]||obs[,{7}];

X2=Ones||obs[,{9}]||obs[,{10}]||obs[,{11}];

**Specify the initial values for newton raphson iteration**

b={2.21,-0.04,0.04,-0.07,2.17,0.01,-0.04,0.03};

max_iter = 10;

n_iter = 1;

diff = 1;
```

```
tolerance = 1e-8;
**Parameter estimates from newton raphson iteration**
do while (( diff > tolerance)& (n_iter<max_iter)) ;
b1=b[{1 2 3 4}];b2=b[{5 6 7 8}];
EY1=(exp(X1*b1)+exp(X1*b1+X2*b2))/(exp(X1*b1)+exp(X2*b2)+exp(X1*b1+X2*b2
));
EY2=(exp(X2*b2)+exp(X1*b1+X2*b2))/(exp(X1*b1)+exp(X2*b2)+exp(X1*b1+X2*b2
));
VY1=(1-exp(X1*b1)/(1+exp(X1*b1))/
   (1-1/((1+exp(X1*b1))#(1+exp(X2*b2))))#
   (exp(X1*b1)/(1+exp(X1*b1))/
   (1-1/((1+exp(X1*b1))#(1+exp(X2*b2)))));
VY2=(1-exp(X2*b2)/(1+exp(X2*b2))/
   (1-1/((1+exp(X1*b1))#(1+exp(X2*b2))))#
   (exp(X2*b2)/(1+exp(X2*b2))/
   (1-1/((1+exp(X1*b1))#(1+exp(X2*b2)))));
Cov_Y1Y2=(1-1/(1-1/((1+exp(X1*b1))#(1+exp(X2*b2))))#
   (exp(X1*b1+X2*b2)/((1+exp(X1*b1))#(1+exp(X2*b2)))/
   (1-1/((1+exp(X1*b1))#(1+exp(X2*b2)))));
EY=EY1//EY2;
Y=Y1//Y2;
X=block(X1`,X2`);
W=(Diag(VY1)||Diag(Cov_Y1Y2))//(Diag(Cov_Y1Y2)||Diag(VY2));
```

```
Cov_T=(X*W*X`);

Score_T=X*(Y-EY);

Pre_b=b;

b=pre_b+solve(Cov_T,Score_T);

diff = sqrt(sum((b-Pre_b)##2));

n_iter = n_iter + 1;

end;

**Estimate probabiliy of being missed by both data sources**

phi=1-1/((1+ exp(X1*b1))# (1+ exp(X2*b2)));

**MLRM estimator for the estimated population size**

ML3_N[j,]=sum(1/phi) ;

ML3_Theta[j,] = b `;

ML3_beta01_est[j,] = b1[1];

ML3_beta11_est[j,] = b1[2];

ML3_beta21_est[j,] = b1[3];

ML3_beta31_est[j,] = b1[4];

ML3_beta02_est[j,] = b2[1];

ML3_beta12_est[j,] = b2[2];

ML3_beta22_est[j,] = b2[3];

ML3_beta32_est[j,] = b2[4];

**MLRM estimator for the variance of estimated population size**

VB=Inv(X*W*X`);

Thi1=exp(X1*b1)#(1+exp(X2*b2))/(exp(X1*b1)+exp(X2*b2)+exp(X1*b1+X2*b2))##2;
```

```
Thi2=exp(X2*b2)#(1+exp(X1*b1))/(exp(X1*b1)+exp(X2*b2)+exp(X1*b1+X2*b2))##2;
Thi=Thi1//Thi2;
V2=(Thi`*X`)*VB*(Thi`*X`)`;
V3=sum((1-phi)/(phi##2));
V0=V2+V3;
alpha=0.05;
**MLRM estimator for measures of performance for the estimated population size**
Bound_ML3[j,1]=sqrt(V0)*(-Probit(1-alpha/2))+ ML3_N[j,1];
Bound_ML3[j,2]=sqrt(V0)*(Probit(1-alpha/2))+ ML3_N[j,1];
Coverage_ML3[j,]=( Bound_ML3[j,1]<=N_True[j,])#( Bound_ML3[j,2]>=N_True[j,]);
WCI_ML3[j,]= Bound_ML3[j,2]-Bound_ML3[j,1];
end ;
**Calculate the average of true diseased population size**
Mean_N=N_True[:,];
**Calculate the average of true diseased population size**
Mean_ML3_N=ML3_N[:,];
**Calculate the average measures of performance for MLRM estimator**
RB_ML3=(Mean_ML3_N-Mean_N)/Mean_N ;
CP_ML3=Coverage_ML3[:,];
Mean_WCI_ML3=WCI_ML3[:,];
Mean_Bound_ML3= Bound_ML3[:,];
RMSE_ML3_N=sqrt(sum((ML3_N-Mean_N)##2)/nsim);
**Measures of performance for parameter estimates**
```

```
RMSE_ML3_beta01=sqrt(sum((ML3_beta01_est-beta01)##2)/nsim);
RMSE_ML3_beta11=sqrt(sum((ML3_beta11_est-beta11)##2)/nsim);
RMSE_ML3_beta21=sqrt(sum((ML3_beta21_est-beta21)##2)/nsim);
RMSE_ML3_beta31=sqrt(sum((ML3_beta31_est-beta31)##2)/nsim);
RMSE_ML3_beta02=sqrt(sum((ML3_beta02_est-beta02)##2)/nsim);
RMSE_ML3_beta12=sqrt(sum((ML3_beta12_est-beta12)##2)/nsim);
RMSE_ML3_beta22=sqrt(sum((ML3_beta22_est-beta22)##2)/nsim);
RMSE_ML3_beta32=sqrt(sum((ML3_beta32_est-beta32)##2)/nsim);
Mean_ML3_Theta = ML3_Theta[:,] ;
**Combine results**
Result_ML3=nsim||correlation||Mean_N||Mean_ML3_N||RB_ML3||CP_ML3||Mean_WC
I_ML3||Mean_Bound_ML3||RMSE_ML3_N||Mean_ML3_Theta||RMSE_ML3_beta01||R
MSE_ML3_beta11||RMSE_ML3_beta21||RMSE_ML3_beta31||RMSE_ML3_beta02||R
MSE_ML3_beta12||RMSE_ML3_beta22||RMSE_ML3_beta32 ;
**Print out results**
Print Result_ML3 ;
quit ;
```