Improving Accuracy of Disease Prevalence Estimates by Combining Information from Administrative Health Records and Electronic Medical Records

by

Saeed Al-Azazi

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirement of the degree of

MASTER OF SCIENCE

Department of Community Health Sciences

University of Manitoba

Winnipeg

ABSTRACT

Introduction: Administrative health records (AHRs) and electronic medical records (EMRs) are the two main sources of population-based data for chronic disease surveillance in Canada. Studies have shown that misclassification errors exist in each data source, which can result in biased estimates of prevalence and incidence. Previous research suggests combining information from data sources, building on their respective strengths to ascertain disease cases.

Purpose & Objectives: The research purpose was to compare different methods to combine information from two error-prone data sources for ascertaining chronic disease cases. The objectives were to: (1) evaluate the bias and precision of several rule-based and probabilistic-based methods using computer simulation, and (2) demonstrate how to apply and use these methods with a numeric example for hypertension case ascertainment.

Methods: Four data-combining methods were compared: (a) rule-based 'OR' method, (b) rule-based 'AND' method, (c) rule-based sensitivity-specificity adjusted (RSSA) method and (c) probabilistic-based sensitivity-specificity adjusted (PSSA) method. The following simulation parameters were investigated: true population prevalence, error-prone data source prevalence, correlation between data sources, number of markers for PSSA method, average correlation amongst markers, and correlation pattern. Relative bias (RB) and mean squared error (MSE) were used for method comparisons. The methods were demonstrated using linked AHRs and EMRs from fiscal years 2005/2006 to 2008/2009 to ascertain cases of hypertension.

Results: The 'OR' method had the lowest RB and MSE when the true prevalence was low, and the RSSA method had the lowest RB and MSE when true prevalence was high. As the correlation between data sources increased, the 'OR' method had the lowest RB and MSE. When the true prevalence was high, correlation between data sources was high and average correlation amongst markers was low, the PSSA method had the lowest RB and MSE. Our numeric results showed a strong correlation between AHRs and EMRs. The estimated prevalence of hypertension from all methods was higher than the estimates using AHRs and EMRs, except for the ‘AND’ method.

Conclusion: The results suggest that no single, optimal data-combining method exists. The ‘OR’ and ‘AND’ methods are influenced by the correlation amongst the data sources, while the RSSA method is dependent on the availability of accurate sensitivity and specificity estimates. The PSSA method performs well only when the true prevalence and correlation amongst data sources are high and average correlations amongst markers is low.

ACKNOWLEDGEMENTS

I would like to first gratefully and sincerely thank my advisor, Dr. Lisa Lix. Thank you for your advice and encouragement during my graduate studies and for all the opportunities you gave me to extend my research knowledge and experience. I am grateful for the financial support provided from the Canadian Institutes of Health Research (CIHR) and the Faculty of Graduate Student (FGS) funding program. I also appreciate the work space at the George and Fay Yee Centre for Healthcare Innovation.

I would also like to express my sincerest thanks to the members of my committee, Dr. Alexander Singer and Dr. Rasheda Rabbani, for providing me with constructive feedback on my thesis. I appreciate their insightful comments and suggestions on my work.

My thanks are extended to the professors in the Department of Community Health Sciences at the University of Manitoba who have taught me during my graduate studies and to those friends of mine for their inputs and valuable discussions.

Finally, I am deeply grateful to my family who has always been devoted and supportive in my quest to achieving my academic goals. It was under their watchful eyes that I gained so much drive to tackle challenges.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	**Definition**
AHR	Administrative Health Record
ATC	Anatomical Therapeutic Chemical
CCDSS	Canadian Chronic Disease Surveillance System
CCS	Charlson comorbidity score
CHD	Coronary heart disease
CHF	Congestive heart failure
CI	Confidence interval
COPD	Chronic obstructive pulmonary disease
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
DIC	Deviance information criterion
DIN	Drug Identification Number
DPIN	Drug Program Information Network
EMR	Electronic Medical Record
FN	False negative
FP	False positive
ICD-9	International Classification of Diseases, 9th revision
ICD-10	International Classification of Diseases, 10th revision
MaPCReN	Manitoba Primary Care Research Network
MCHP	Manitoba Centre for Health Policy
MI	Multiple Imputation
MSE	Mean square error
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
PHAC	Public Health Agency of Canada
PHIN	Personal health identification number
PPV	Positive predictive value
PSRF	Potential scale reduction factors
PSSA	Probabilistic-based sensitivity-specificity adjusted
RB	Relative Bias
RD	Renal disease
RSSA	Rule-based sensitivity-specificity adjusted
SA	Substance abuse
SARD	Systemic autoimmune rheumatic disease
TN	True negative
TP	True positive

CHAPTER 1 – INTRODUCTION

1.1 Background

Administrative health records (AHRs) and electronic medical records (EMRs) are the two main sources of population-based data for chronic disease surveillance in Canada. AHRs, including hospital discharge abstract, physician billing claims and prescription drug records, are currently used in Canadian provinces and territories through the Public Health Agency of Canada's (PHAC's) Canadian Chronic Disease Surveillance System (CCDSS) (Pelletier et al., 2012; Dai et al., 2013; O'Donnell, 2013; Robitaille et al., 2013). The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) is the only pan-Canadian EMR data source for chronic disease surveillance (Coleman et al., 2015; Williamson et al., 2014). Established in 2008, the CPCSSN extracts and processes EMR data from 10 primary care research networks across Canada (Garies et al., 2017).

Chronic disease case ascertainment algorithms, the criteria used to ascertain individuals with a specific condition, have been developed independently for AHRs and EMRs. In EMRs, the components of a case ascertainment algorithm include combinations of diagnosis and prescription drug codes and structured and unstructured (i.e., text) information drawn from a number of sections, including laboratory test results (Williamson et al., 2014). In AHRs, the components of a case ascertainment algorithm include the type of data source, diagnostic and prescription drugs codes, number of records with the code(s), and number of years of data (Lix et al., 2006).

Validation studies, in which diagnosed cases are compared with clinically-confirmed cases, have been conducted to assess the accuracy of case ascertainment algorithms for both AHRs and EMRs. These studies have shown that misclassification errors exist in both data sources (Coleman et al. 2015; Kadhim-Saleh et al., 2013; Lix et al., 2006; Quan et al., 2008; Tu et al., 2007; Williamson et al., 2014), including false negative cases in which an individual is incorrectly classified as not having a disease and false positives in which an individual is incorrectly classified as having a disease (Valle et al., 2015). For example, Quan et al. (2008) compared AHRs to data abstracted from physician charts to assess the accuracy of case ascertainment algorithms of AHRs for identifying individuals with hypertension. The sensitivity (i.e., the proportion of correctly identified positive cases) and specificity (i.e., the proportion of correctly identified negative cases) of the case ascertainment algorithms of hypertension from AHRs were 75% and 94%, respectively. Williamson et al. (2014) found that the sensitivity and specificity of the case ascertainment algorithms of hypertension from EMRs were 85% and 94%, respectively. These results indicate that both data sources are imperfect for case ascertainment, which can result in biased estimates of disease prevalence and incidence. In addition, inconsistencies may exist between the two sources; individuals who are identified as disease cases in AHRs may not be disease cases in EMRs, resulting in different disease estimates from the two sources (Atwood et al., 2013). Each source has strengths and limitations (Birtwhistle & Williamson, 2015; Quan et al., 2012; Singer et al., 2016) as described in Table 1.1.

Table 1.1: Strengths and limitations of using administrative health records (AHRs) and Canadian Primary Care Sentinel Surveillance Network (CPCSSN) electronic medical records (EMRs) for research and surveillance

Source	Limitations	Strengths
AHRs	Intended for provider reimbursement and health system management rather than for research and surveillance	Complete population coverage
	Do not contain patient clinical information, such as laboratory test results or biological measurements	Available for many years
	A single diagnosis is recorded in physician claims for many provinces (e.g., Manitoba)	Include records of all filled prescriptions and dates of fills/refills
CPCSSN EMRs	Were developed to be used for patient care rather than for research or surveillance	Include laboratory results, which may provide clinical indications of chronic diseases
	Contain unstructured data, which can be challenging to use for research and surveillance	Contain health condition lists to record medical history of patients
	CPCSSN is a voluntary network of primary care providers; the patients who see these providers may not be fully representative of the Canadian outpatient population	Contain some health risk factor information, such as self-reported smoking and alcohol use, which is important for disease risk prediction

Previous research has suggested combining information from two or more data sources, to build on the strengths of each source when ascertaining disease cases. One of these sources may contain no misclassification error, or both sources may contain misclassification error. For the former, a validation sample, in which disease status is known without error, may be used to correct for misclassification bias in an error-prone data source (He & Zaslavsky, 2009; Yucel &

Zaslavsky, 2005; Zheng et al., 2006). For the latter, possible solutions range from simple rule-based to probabilistic-based and latent-class modelling approaches, which rely on estimates of sensitivity and specificity of disease case ascertainment algorithms to correct for misclassification bias (Bernatsky et al., 2005; Dendukuri & Joseph, 2001; He et al., 2014; Reitsma et al., 2009). Rule-based methods classify individuals as having the disease of interest based on pre-specified data-combining rules (Alonzo & Pepe, 1999; Martin et al., 2004; Schiller et al., 2016). Adjusted rule-based methods ascertain the number of disease cases by using the accuracy of case ascertainment algorithms as weights (Couris et al., 2002; Couris et al., 2009; Hadgu et al., 2005). Probabilistic-based methods fit regression models, including Bayesian models, to select values of the true disease status for a pre-specified distribution. These models sample from the posterior distribution conditional on disease markers, using an iterative process (He et al., 2014).

Data-combining methods based on both deterministic and probabilistic models have been proposed but few, if any, studies have compared these methods. As well, there has been limited investigation about the factors that may influence the accuracy of these methods. Such investigations could assist researchers to make informed decisions when choosing methods to combine data sources for case ascertainment and provide recommendation for ease of use. Accurate information about disease prevalence and incidence is needed for understanding disease burden in the population, and for planning disease treatment and management programs.

1.2 Purpose and objectives

The purpose of this research was to compare different methods to combine information from two error-prone data sources for ascertaining chronic disease cases. The specific objectives were:

1. To evaluate the bias and precision of several rule-based and probabilistic-based methods using computer simulation; and
2. To demonstrate how to apply and use these methods with a numeric example for hypertension case ascertainment.

1.3 Thesis organization

This thesis focuses on the use of data-combining methods for chronic disease surveillance. Chapter 2 presents relevant literature on the following topics: a) case ascertainment with one error-prone data source, b) case ascertainment with one error-prone data source and one error-free data source, c) case ascertainment with two or more error-prone data sources, and d) accuracy of case ascertainment algorithms for chronic diseases in AHRs and EMRs. Chapter 3 contains the methods for the simulation study and the numeric example. Chapter 4 presents the results of the simulation study. Chapter 5 presents the results of the numeric example. The last chapter includes the study discussion, conclusions, and directions for future research.

CHAPTER 2 – LITERATURE REVIEW

This literature review encompasses the following topics: case ascertainment with one error-prone data source, case ascertainment with one error-prone data source and one error-free data source, case ascertainment with two or more error-prone data sources, and accuracy of case ascertainment algorithms for chronic diseases in AHRs and EMRs.

2.1 Case ascertainment of one error-prone data source

First, we define the sensitivity and specificity for case ascertainment algorithms of an error-prone data source. Consider a two-way table of observed disease status from an error-prone data source versus disease status from a validation study that contains error-free data (Table 2.1). Let D_i denote a binary indicator of the true disease status for the *i*th individual (*i* = 1,…, *n*), where 1 = disease case and 0 = non-disease case. Let Y_{1i} denote the observed disease status for the *i*th individual from an error-prone data source (e.g., AHRs), where 1 = disease case and 0 = non-disease case.

Table 2.1: Two-way table for ascertaining disease status from an error-prone data source (Y_1) and a data source containing true disease status (D)

		$\boldsymbol{D}$		
		1	**0**	**Total**
$\boldsymbol{Y_1}$	**1**	TP	FP	TP+FP
	0	FN	TN	FN+TN
Total		TP+FN	FP+TN	TP+FP+FN+TN

The diagonal cells represent correctly classified cases known as true positives (TP) and true negatives (TN), respectively. The off-diagonal cells indicate misclassified cases known as false positives (FP) and false negatives (FN), respectively. Sensitivity is estimated as the number of true positive cases divided by the total number of true disease cases (TP+FN), defined in terms of probability as P(Y_1=1 | D=1). Specificity is estimated as the number of true negative cases divided by the total number of true non-disease cases (FP+TN), defined in terms of probability as P(Y_1=0 | D=0).

A data source is said to be imperfect when sensitivity and/or specificity of its case ascertainment algorithms do not equal 1.00. From Table 2.1, if sensitivity is less than 1.00, then FNs will occur. That is, individuals who are true disease cases will be classified by the error-prone data source as non-disease cases. If specificity is less than 1.00, FPs will occur where individuals who are true non-disease cases will be classified by the error-prone data source as disease cases. If more individuals are classified as FNs than FPs, then the overall probability of the disease status will be underestimated. If the numbers of FNs and FPs are not equal to zero, estimates of disease prevalence and incidence will be biased.

To illustrate the effect of misclassification bias on disease prevalence, let $prev_{\mathrm{T}}$ denote the true disease prevalence and $prev_{Y_1}$ denote the observed disease prevalence from AHRs. Suppose we wish to evaluate a population where the true prevalence, $prev_{\mathrm{T}}$= 0.30. If AHRs has less-than-perfect sensitivity of 0.70, and perfect specificity of 1.00, then by the laws of probability (Rogan & Gladen, 1978): $prev_{Y_1}$= $prev_{\mathrm{T}}$ (S_n) + (1.00 – $prev_{\mathrm{T}}$) (1.00 – S_p) = 0.30 x 0.70 + (1.00 – 0.30) x (1.00 –1.00) = 0.21. Hence, the observed disease prevalence from AHRs will underestimate the true prevalence. Otherwise, if AHRs have less-than-perfect specificity of 0.80, and perfect sensitivity of 1.00, then $prev_{Y_1}$= 0.30 x 1.00 + (1.00 – 0.30) x (1.00 – 0.80) = 0.44. In this case, the observed disease prevalence from AHRs will overestimate the true prevalence. When both sensitivity and specificity are less than 1.00, combining both examples, $prev_{Y_1}$= 0.30 x 0.70 + (1.00 – 0.30) x (1.00 – 0.80) = 0.35. Thus, accurate estimates of disease prevalence cannot generally be obtained from an error-prone data source.

Underestimation or overestimation of disease prevalence will result from the less-than-perfect sensitivity and specificity of case ascertainment algorithms for a data source. In general, validation studies done using AHRs and EMRs have found that specificity is higher than sensitivity (Coleman et al. 2015; Lix et al., 2006; Quan et al., 2008; Williamson et al., 2014). As illustrated above, high specificity and low sensitivity will result in under-reporting of the total number of disease cases.

2.2 Case ascertainment with one error-prone data source and one error-free data source

Research in the area of combining information from two sources has looked into the situation where one data source is assumed to be accurate and complete for the variable of interest in a validation sample of the population, while the other source is error-prone, that is, the source is subject to misclassification or missing data (He et al., 2009; Raghunathan et al., 2007; Schenker et al., 2010; Schenker et al., 2007; Yucel & Zaslavsky, 2005; Zheng et al., 2006). Misclassification in the error-prone data source can be corrected using multiple imputation (MI) methods (Rubin 1987). These methods fill in missing observations several times to create multiple completed datasets. The results obtained from separate sets of complete data are combined into a single inference using simple rules (Rubin 1987). MI methods replace each missing value with a set of plausible values that represent the uncertainty about the correct value to impute.

Raghunathan et al. (2007) and Schenker et al. (2010) discussed combining information from the National Health Interview Survey (NHIS), which has self-reported health data, and the National Health and Nutrition Examination Survey (NHANES), which has both self-reported health data and clinical measures. They constructed a reporting error model that predicts an outcome given the self-reported data and covariates. In their context, they obtained a validation sample, which is a sample of individuals for whom both the true value of interest and reported value of interest are observed (Pepe, 1992). Comparable reporting error models were proposed by other researchers who obtained similar validation samples to correct for misclassification bias (He & Zaslavksy, 2009; Yucel & Zaslavsky, 2005).

Yucel and Zaslavsky (2005) considered imputation methods for misclassified binary treatment variables in AHRs. The study used AHRs from a cancer registry that collected data on treatment and survival for all incident colorectal cancer cases in California. The researchers used reported chemotherapy treatment status in which the true chemotherapy status was obtained by surveying physicians or reviewing office records (i.e., validation sample). Corrected treatment status was obtained by modelling the reporting processes and the relationship of the true treatment status to other markers in the registry. He and Zaslavksy (2009) extended this work to allow for multivariate responses.

He and Zaslavksy (2009) considered under-reporting of adjuvant chemotherapy and radiation therapy for breast cancer in a cancer registry. They imputed the treatment information from a single inaccurately-reported therapy variable using a model based on a validation sample. Using an iterative process, known as the Gibbs sampling algorithm, they estimated the model parameters and imputed treatment status. Specifically, the model imputed the adjuvant chemotherapy and radiation therapy data separately for each patient for uncollected true therapy status in the registry sample.

2.3 Case ascertainment with two or more error-prone data sources

Combining two or more error-prone data sources may correct for misclassification bias when ascertaining disease cases (Bernatsky et al., 2005; Bernatsky et al., 2011; Dendukuri & Joseph, 2001; He et al., 2014; Joseph et al., 1995). Methods that do not use a validation sample include rule-based methods, Frequentist models, and Bayesian models (Bernatsky et al., 2005; Cook et al., 2017; He et al., 2014).

Rule-based methods were proposed by Alonzo and Pepe (1999). A simple rule-based method is one in which two or more case ascertainment algorithms are applied to the data for all patients and a pre-specified (i.e., deterministic) rule is then used to classify patients as having the target disease or not (Martin et al., 2004; Schiller et al., 2016). Several definitions for the presence of the target disease can be used; the choice depends on whether emphasis will be given to detecting or excluding the target disease and the characteristics of the available case ascertainment algorithms (Reitsma et al., 2009). For example, the “OR” method identifies a patient to be a disease case if he/she was identified as a disease case by either one of the case ascertainment algorithms; whereas, the “AND” method identifies a patient to be a disease case if he/she was identified as a disease case by two or more two case ascertainment algorithms.

The simplest rule-based methods give equal weight to all algorithms. Rule-based methods that incorporate weighting factors have also been proposed (Hadgu et al., 2005). Weighted methods use information about the accuracy of case ascertainment algorithms (i.e., sensitivity and specificity) from published literature as weights (Couris et al., 2002; Couris et al., 2009).

Probabilistic models based on the Bayesian framework have also been proposed. The data, through the likelihood function, are combined using prior information to derive posterior distributions of the values of interest using Bayes' theorem. These posterior distributions contain updated beliefs about the values of interest, after taking into account the information provided by the data (Joseph et al., 1995).

Bernatsky et al. (2011) proposed Bayesian latent class models to address under-ascertainment of cases of systemic autoimmune rheumatic diseases (SARDs) in AHRs. Their method identifies disease status using error-prone data sources and prior information about the sensitivity and specificity of each of the error-prone data sources. Bernatsky et al. selected prior information for the Bayesian model based on previous research. Estimates of SARDs prevalence from the Bayesian latent class models were consistent with estimates from other population-based data sources, including health surveys.

He et al. (2014) considered the problem where the true outcome of interest is missing (i.e., latent) for all individuals. Specifically, the researchers identified two data sources that contained reported disease status; data in both sources were subject to misclassification and missingness. The researchers decomposed the joint model of the true outcome and the two reported statuses into an "outcome" model (true outcome given disease markers) and a "reporting" model (reported status given true outcome and disease markers). Their reporting model relied on a few assumptions. First, it was assumed that the two reported statuses were conditionally independent. In other words, the joint distribution of the two reported statuses was conditional on the true disease status and observed markers. Given this independence, the joint distribution was decomposed by multiplying the probability of first and second reported statuses. Second, it was assumed that the two data sources were subject to under-reporting only. That is, the reported status from each data source had less-than-perfect sensitivity but perfect specificity.

Cook et al. (2017) proposed a maximum likelihood estimator that corrects for misclassification in data from multiple sources. They considered the problem of under-reporting or incompleteness of data from primary and secondary media outlets for events such as strikes, protests, and conflicts. Their estimator allows researchers to estimate the extent of misclassification in a data source, and to obtain corrected estimates of the event of interest. Generalization of their estimator allows for models where misclassification is dependent upon covariates. Similar to the assumptions made by He et al. (2014), the data sources in their context were assumed to have less-than-perfect sensitivity but perfect specificity.

2.4 Accuracy of case ascertainment algorithms for chronic diseases in administrative health records and electronic medical records

A large number of studies have investigated the validity of case ascertainment algorithms in AHRs and EMRs for ascertaining cases of chronic disease (Coleman et al. 2015; Kadhim-Saleh et al., 2013; Lix et al., 2006; Quan et al., 2008; Tu et al., 2007; Williamson et al., 2014). Data sources used as the reference (i.e., validation or error-free) data source include medical records, patient surveys, and clinical laboratory test results. This section will focus primarily on hypertension, because it is the focus of the numeric example that will be considered in this study.

Quan et al. (2008) compared data in AHRs to data abstracted from physician charts to assess the accuracy of AHR case ascertainment algorithms for hypertension. Physician charts were randomly selected for rural and urban areas from the Canadian provinces of Alberta and British Columbia during the years 2001 and 2004. The sensitivity and specificity of the hypertension case ascertainment algorithms for AHRs were 75% and 94%, respectively.

Williamson et al. (2014) validated case ascertainment algorithms used to identify eight common chronic conditions in primary care EMRs: chronic obstructive pulmonary disease (COPD), dementia, depression, diabetes, hypertension, osteoarthritis, Parkinson’s disease, and epilepsy. Patient’s charts were reviewed by research assistants and residents who were blinded to the diagnosis. The sensitivity and specificity of the hypertension case ascertainment algorithms for EMRs were 85% and 94%, respectively. For the other conditions, sensitivity ranged from 78% for osteoarthritis to 95% for diabetes, Parkinson’s disease and epilepsy; whereas, specificity was above 94% for all conditions.

Several studies have examined the accuracy of diagnosis codes for AHR and EMR case ascertainment for various chronic conditions. In general, these validation studies have demonstrated high specificity but low sensitivity of diagnoses, although sensitivity and specificity vary across chronic diseases (Coleman et al. 2015; Kadhim-Saleh et al., 2013; Lix et al., 2006; Quan et al., 2008; Tu et al., 2007; Williamson et al., 2014).

CHAPTER 3 – METHODS

This study was conducted using computer simulation techniques and analyses involving a numeric example based on Manitoba's AHRs and EMRs. This chapter describes bias in prevalence estimates with two error-prone data sources and data-combining methods for two or more error-prone data sources. For the simulation study, we describe the simulation parameter values, data generation steps, and performance measures. For the numeric example, we define the data sources, the study cohort and case ascertainment markers.

3.1 Bias in prevalence estimates with two error-prone data sources

Consider data sources Y_j, j = 1, 2, that capture disease status for n individuals. Denote the true disease status for the i^{th} individual (i = 1, 2,…, n), by D_i, which takes a value of 1 when the individual has the disease of interest, and 0 when the individual does not have the disease of interest. Denote the observed disease status from the j^{th} data source for the i^{th} individual by Y_{ij} , which takes a value 1 when the individual is classified as a disease case by data source j, and 0 when the individual is classified as a non-disease case by source j. Sensitivity of observed disease status for the j^{th} data source is denoted as Sn_j= P(Y_j= 1 | D= 1) , and specificity is denoted as Sp_j= P(Y_j= 0 | D= 0).

When sensitivity and/or specificity for both data sources are less than 1, the data sources are imperfect and observed disease status from each data source will be biased. To illustrate, a two-way contingency table of observed disease status from AHRs (Y_1) and EMRs (Y_2) is provided in Table 3.1. Let Y_{i1} and Y_{i2} denote the observed disease status for the i^{th} individual from AHRs and EMRs, respectively. Let $prev_{Y_1}$ and $prev_{Y_2}$ denote the naïve disease prevalence estimates from AHRs and EMRs, respectively. The naïve disease prevalence estimates are:

$$prev_{Y_1} = \frac{n_{11}+n_{10}}{n_{11}+n_{01}+n_{10}+n_{00}} \qquad (3\text{-}1)$$

$$prev_{Y_2} = \frac{n_{11}+n_{01}}{n_{11}+n_{01}+ n_{10}+ n_{00}} \tag{3-2}$$

where n_{11} is the number of individuals classified as disease cases in both data sources, n_{10} is the number of individuals classified as disease cases in AHRs but not EMRs, n_{01} is the number of individuals classified as disease cases in EMRs but not AHRs, and n_{00} is the number of individuals not classified as disease cases in either data source. These naïve estimates do not take into account the likelihood that some individuals could be incorrectly classified, that is, that FPs and/or FNs may exist in the data.

Table 3.1: Two-way table of observed disease status from AHRs (Y_1) and EMRs (Y_2).

		Y_1		
		1 (Disease Present)	**0 (Disease Absent)**	**Total**
Y_2	**1 (Disease Present)**	n_{11}	n_{01}	$n_{11}+n_{01}$
	0 (Disease Absent)	n_{10}	n_{00}	$n_{10}+n_{00}$
Total		$n_{11}+n_{10}$	$n_{01}+n_{00}$	$n_{11}+n_{01}+ n_{10}+ n_{00}$

3.2 Data-combining methods for two or more error-prone data sources

In this section, methods that combine information from error-prone data sources to ascertain cases will be defined, including: (a) rule-based ‘OR’ and ‘AND’ methods (Alonzo & Pepe, 1999), (b) rule-based sensitivity-specificity adjusted (RSSA) method (Couris et al., 2002), and (c) probabilistic-based sensitivity-specificity adjusted (PSSA) method (He et al., 2014).

3.2.1 Rule-based ‘OR’ and ‘AND’ methods

Rule-based ‘OR’ and ‘AND’ methods use a pre-specified rule to classify individuals as having the target disease or not. From Table 3.1, the ‘OR’ method assigns a ‘1’ for an individual if he/she was classified as a ‘1’ in either data source. The ‘AND’ method assigns a ‘1’ for an individual only if he/she was classified as a '1' in both data sources. Each method produces a different estimate of the number of disease and non-disease cases.

Let $prev_{OR}$ and $prev_{AND}$ represent the estimated disease prevalence using the 'OR' and 'AND' methods, respectively. From Table 3.1, prevalence is estimated as:

$$prev_{OR} = \frac{n_{11} + n_{01} + n_{10}}{n_{11} + n_{01} + n_{10} + n_{00}} \tag{3-3}$$

$$prev_{AND} = \frac{n_{11}}{n_{11} + n_{01} + n_{10} + n_{00}} \tag{3-4}$$

where $n_{11}, n_{01}, n_{10}, n_{00}$ are defined in equations (3-1) and (3-2).

The 'OR' method classifies more individuals as disease cases than either data source alone, making it suitable when the sensitivities of the two data sources are low, that is, when both data sources tend to capture true disease cases poorly. On the other hand, if both data sources tend to poorly capture true non-disease cases, then the 'AND' method is preferable (Schiller et al., 2016).

The main assumptions made by the 'OR' and 'AND' methods include: (i) 'OR' and 'AND' methods treat observed disease status from each data source as 100% sensitive and specific, and (ii) observed disease status from the two data sources is assumed to be conditionally independent. The second assumption implies that for a given individual, observed disease status from one data source has no impact on observed disease status of the other data source. Further, these methods ignore the possible (and maybe important) relationship between true disease status and other variables. The next method relaxes the first assumption.

3.2.2 Rule-based sensitivity-specificity adjusted (RSSA) method

The RSSA method uses information about the accuracy of case ascertainment algorithms from prior validation studies to adjust the estimates of the number of true disease cases. That is, sensitivity and specificity of the case ascertainment algorithms for a prior data source are used to correct for misclassification bias (Couris et al., 2002; Couris et al., 2009). Let π= P(D_i= 1) be the probability of the true disease status for the ith individual. The joint probability of the observed disease status from the two error-prone data sources, (Y_1= y_1, Y_2= y_2), where y_1, y_2= 0, 1, is defined as:

$$\mathrm{P}(Y_1 = y_1, Y_2 = y_2) = \mathrm{P}(Y_1 = y_1, Y_2 = y_2 | D = 1)\ \mathrm{P}(\mathrm{D} = 1) + \mathrm{P}(Y_1 = y_1, Y_2 = y_2 | D = 0)\ \mathrm{P}(D = 0)$$

$$= \mathrm{P}(Y_1 = y_1, Y_2 = y_2 | D = 1)\ \pi + \ \mathrm{P}(Y_1 = y_1, Y_2 = y_2 | D = 0)\ (1 - \pi) \qquad \text{(3-5)}$$

The RSSA method assumes that the observed disease status from the two data sources is conditionally independent given the true disease status. From equation (3-5), the joint probabilities are equal to the product of the marginal probabilities:

$$\mathrm{P}(Y_1 = y_1 | D = 1)\ \mathrm{P}(Y_2 = y_2\ | D = 1)\ \pi + \mathrm{P}(Y_1 = y_1 | D = 0)\ \mathrm{P}(Y_2 = y_2\ | D = 0)\ (1 - \pi) \qquad \text{(3-6)}$$

Given this conditional independence, the number of individuals classified as having the disease of interest in both data sources, n_{11}, and the number of individuals classified as not having the disease of interest in both data sources, n_{00}, are assumed to be correct. Therefore, adjustments are only made to the total number of individuals with discordant disease status, that is (Y_1, Y_2) = (1, 0) and (0, 1). From Table 3.1, the number of individuals in the discordant cells (i.e., n_{10} and n_{01}) is allocated, in part, to either of the two cells containing correctly classified individuals (i.e., n_{11} and n_{00}) (Naaktgeboren et al., 2013).

Consider our example of hypertension. Values of sensitivity and specificity of AHR and EMR case ascertainment algorithms of hypertension can be used to weight the disconcordant cells. For AHRs, we considered three Canadian validation studies about hypertension (Lix et al., 2006; Quan et al., 2008; Tu et al., 2007) that were previously identified in a systematic review (Pace et al., 2017). For EMRs, we identified the only three Canadian validation studies about hypertension done to date (Coleman et al. 2015; Kadhim-Saleh et al., 2013; Williamson et al., 2014). The average values of sensitivity and specificity identified from Canadian validation studies about hypertension were 0.72 and 0.95 for AHRs and 0.87 and 0.90 for EMRs. The two discordant cells, when Y_1 and Y_2 takes on values (1, 0) and (0, 1), are corrected using these sensitivity and specificity values as weights.

Weight (D= 1) = P(D= 1 | Y_1= 1, Y_2= 0) + P(D= 1 | Y_1= 0, Y_2= 1)

$$= \text{P}(Y_1{=}\,1 \mid D{=}\,1)\ \text{P}(Y_2{=}\,0 \mid D{=}\,1)\ n_{10} + \text{P}(Y_1{=}\,0 \mid D{=}\,1)\ \text{P}(Y_2{=}\,1 \mid D{=}\,1)\ n_{01}$$

$$= (Sn_1)\ (1 - Sn_2)\ n_{10} + (1 - Sn_1)\ (Sn_2)\ n_{01} \qquad (3\text{-}7)$$

Weight (D = 0) = P(D = 0 | Y_1 = 1, Y_2 = 0) + P(D = 0 | Y_1 = 0, Y_2 = 1)

$$= \text{P}(Y_1 = 1 \mid D = 0)\ \text{P}(Y_2 = 0 \mid D = 0)\ n_{10} + \text{P}(Y_1 = 0 \mid D = 0)\ \text{P}(Y_2 = 1 \mid D = 0)\ n_{10}$$

$$= (1 - Sp_1)\ (Sp_2)\ n_{10} + (Sp_1)\ (1 - Sp_2)\ n_{10} \qquad (3\text{-}8)$$

The RSSA method, too, ignores the possible associations between the true disease status and other variables. The method considered next incorporates this association.

3.2.3 Probabilistic-based sensitivity-specificity adjusted (PSSA) method

The PSSA method recognizes the possible associations between true disease status and other variables, known as disease markers (He et al., 2014). Here, the sensitivities and specificities of the two data sources are related by marker X_i, for the i^{th} individual. In addition to the previous notation, let sensitivity and specificity of the j^{th} data source in individuals with the vector of markers, **X**, be written as $Sn_j(x)$ = P(Y_j = 1 | D = 1, **X**) and $Sp_j(x)$ = P(Y_j = 0 | D = 0, **X**), j = 1, 2. These ascertainment accuracy parameters are modelled via a Bayesian regression model with a probit link function, which is used to link the outcome variable, D, to the vector of markers **X**. Let θ indicate some parameters governing the process of ascertainment. The PSSA method models the joint distribution of the data as:

$$\text{P}(Y_1, Y_2, D \mid \mathbf{X}, \theta) = \text{P}(D \mid \mathbf{X}, \theta_D)\ \text{P}(Y_1, Y_2 \mid D, \mathbf{X}, \theta_Y) \qquad (3\text{-}9)$$

The first term named the ‘outcome’ model, relates the true disease status to marker vector **X**, with regression parameters θ_D. The second term named the ‘reporting’ model (or the sensitivity-specificity model), characterizes the observed disease status from two data sources given true disease status, markers, and parameter θ_Y. The PSSA method assumes that the observed disease status from the two data source is independent, conditional on the true disease status and markers, that is,

$$\mathrm{P}(Y_1, Y_2 \mid D, \mathbf{X}, \theta_Y) = \mathrm{P}(Y_1 \mid D, \mathbf{X}, \theta_{Y1})\,\mathrm{P}(Y_2 \mid D, \mathbf{X}, \theta_{Y2}) \tag{3-10}$$

The complete data model is given by:

$$Z_{Di} = \Phi^{-1}[\mathrm{P}(D_i = 1 \mid X_i)] = \mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{D}} \tag{3-11}$$

$$Z_{R1i} = \Phi^{-1}[\mathrm{P}(Y_{1i} = 1 \mid D_i = 1, X_i)] = \mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R1}} \tag{3-12}$$

$$Z_{R2i} = \Phi^{-1}[\mathrm{P}(Y_{2i} = 0 \mid D_i = 0, X_i)] = \mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R2}} \tag{3-13}$$

where Z_{D_i}, Z_{R1i}, and Z_{R2i} are normally distributed latent (i.e., unobserved) variables, $\boldsymbol{\beta}_{\mathrm{D}}$, $\boldsymbol{\beta}_{\mathrm{R1}}$, and $\boldsymbol{\beta}_{\mathrm{R2}}$ are vectors of fixed-effects parameters, $\mathbf{X}_i^T$ is the vector of markers, $^{\mathrm{T}}$ denotes the transpose operator, and Φ^{-1} is the probit link function.

The PSSA method uses a posterior sampling procedure known as data augmentation (DA) (Tanner & Wong, 1987) to draw values of the unobserved true disease status *D* by sampling from the posterior distribution, given markers **X**. Specifically, the Gibbs sampling algorithm, which is an iterative Markov chain Monte Carlo (MCMC) technique, is used to draw values of the unobserved true disease status *D* (Casella & George, 1992). Following multiple iterations of the algorithm, the generated draws of *D* eventually converge to the stationary distribution, which is the desired posterior distribution. The DA procedure is implemented in two steps:

Step 1 (Imputation step): Draw *D* from a Bernoulli distribution with the conditional probability of unobserved true disease status, *D*, given the observed disease status from the *j*th data source. Do this when Y_1 and Y_2 take on values (1, 1), (1, 0), (0, 1) and (0, 0). For example, the decomposition of the joint distribution of P(*D*, Y_1, Y_2, $\mathbf{X}$) when $Y_1$and Y_2 takes on values (0,0) is calculated as:

P(*D* = 1 | Y_1 = 0, Y_2 = 0, $\mathbf{X}$, θ)

$$= \frac{\mathrm{P}(Y_1 = 0 \mid D = 1, \mathbf{X})\, \mathrm{P}(\, Y_2 = 0 \mid D = 1, \mathbf{X})\;\; P(D = 1|\mathbf{X})}{\mathrm{P}(Y_1 = 0 \mid D = 1, \mathbf{X})\, \mathrm{P}(Y_2 = 0 \mid D = 1, \mathbf{X})\, \mathrm{P}(D = 1|\mathbf{X}) +\; \mathrm{P}(Y_1 = 0 \mid D = 0, \mathbf{X})\, \mathrm{P}(\, Y_2 = 0 \mid D = 0, \mathbf{X})\, \mathrm{P}(D = 0|\mathbf{X})}$$

$$= \frac{\Phi\left(-\mathbf{X}_i^T \boldsymbol{\beta}_{\mathrm{R}_1}\right) \Phi\left(-\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R}_2}\right) \Phi\left(\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{D}}\right)}{\Phi\left(-\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R}_1}\right) \Phi\left(-\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R}_2}\right) \Phi\left(\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{D}}\right) + \Phi\left(\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R}_1}\right) \Phi\left(\mathbf{X}_i^T\, \boldsymbol{\beta}_{\mathrm{R}_2}\right) \Phi\left(-x_i^T\, \boldsymbol{\beta}_{\mathrm{D}}\right)} \qquad \text{(3-14)}$$

Step 2 (Posterior step): Draw new values of θ's conditional on the imputed *D* via the Gibbs sampling algorithm for probit models (Chib & Greenberg, 1998). That is, latent variables Z_{D_i}, Z_{R1i}, and Z_{R2i}, and fixed-effects parameters $\boldsymbol{\beta}_{\mathrm{D}}$, $\boldsymbol{\beta}_{\mathrm{R1}}$, and $\boldsymbol{\beta}_{\mathrm{R2}}$ are drawn from truncated and multivariate normal distributions, respectively. Details of the Gibbs sampling algorithm are found in the appendix of He et al. (2014).

As in all Bayesian approaches, prior distributions are required for each parameter, *θ*. Vague priors for the parameters are the easiest to consider; in this study, flat priors for $\boldsymbol{\beta}_{\mathrm{D}}$, $\boldsymbol{\beta}_{\mathrm{R1}}$, and $\boldsymbol{\beta}_{\mathrm{R2}}$ were imposed (Burton, 1994). A flat (or uniform) prior is one that does not favor any particular value for the parameter (Spiegelhalter et al., 2004).

Convergence assessment of the generated draws of *D* determines whether the iterations of the MCMC chain reached the target posterior distribution or longer iterations are needed. In this study, we plotted the results for a visual graphical assessment using trace plots, which are plots of the iteration number against the value of the draw of the parameter(s) of interest. By graphing the trace plot of the chain starting from different starting positions, one can assess if there is convergence to the same posterior distribution. In addition, we used the Gelman-Rubin diagnostic that is based on running and analyzing the difference between two or more chains (Gelman and Rubin, 1992). The Gelman-Rubin diagnostic measures whether there is a significant difference between the variance within several chains and the variance between several chains by a value called “potential scale reduction factors” (PSRF). Large differences between these variances indicate nonconvergence. If the chains have converged to the target posterior distribution, then the PSRF should be close to 1 (Gelman and Rubin, 1992). The *PSRF* is given by:

$$\text{PSRF} = \sqrt{\frac{Var\ (\theta)}{W}}\ , \tag{3-15}$$

where W is the within-chain variance and $Var(\theta) = (1 - \frac{1}{n})\ W + \frac{1}{n} B$ is the estimated variance of the posterior distribution as a weighted average of W and between-chain (B) variance.

3.3 Simulation study

3.3.1 Simulation parameter values

A simulation study was conducted to assess the bias and accuracy of data-combining methods under realistic data-analytic conditions. Observations for the simulation study were drawn from an infinitely large population. The following parameters and data characteristics were investigated: true population prevalence ($prev_{\mathrm{T}}$), error-prone data source prevalence ($prev_{Y_1}, prev_{Y_2}$), correlation between data sources ($\rho_{Y_1Y_2}$), number of markers for PSSA method (N_x), average correlation amongst markers ($\bar{\rho}_x$) and correlation pattern ($\bar{\rho}_{x\ (\text{pattern})}$). Each parameter and its selected values are shown in Table 3.2.

Table 3.2: Simulation study parameters and values

Simulation Parameter	**Values**
$prev_{\mathrm{T}}$	20%, 10%
$prev_{Y_1}$	18%, 15%, 8%, 5%
$prev_{Y_2}$	15%, 10%, 7%, 5%
$\rho_{Y_1Y_2}$	0.65, 0.85
N_x	8, 16
$\bar{\rho}_x$	0.00, 0.20, 0.50
$\bar{\rho}_x$ (pattern)	$\bar{\rho}_x$ (exchangeable), $\bar{\rho}_x$ (unstructured)

When true prevalence in the population was set to 20%, the error-prone data source prevalence was set to (18%, 15%) for source 1 and 2, respectively, with data source correlations of 0.65 and 0.85. Similarly, we set prevalence to (18%, 10%), and (15%, 15%) for source 1 and 2, with data source correlation of 0.65 and 0.85. Data source prevalence and correlations were manipulated by varying the degree of sensitivity and specificity for each data source. The number of markers for the PSSA method was fixed to 8 or 16. Three values of the average correlation amongst markers ($\bar{\rho}_x$) were considered: 0.00, 0.20, and 0.50, with an exchangeable or unstructured correlation pattern amongst the markers. An exchangeable correlation pattern is one where a constant correlation value is imposed amongst all the markers (Barnett et al., 2010). On the other hand, an unstructured correlation pattern allows for variability in the magnitude of correlations amongst all the markers (Barnett et al., 2010). A total of 144 combination of simulation conditions were considered for each of the data-combining methods. A summary of each combination of simulation conditions are shown in Appendix A, Table A.1.

True prevalence of 20% was chosen to reflect the estimated prevalence of hypertension observed in previous studies about prevalence for the entire population (Padwal et al., 2016), whereas the true prevalence of 10% was chosen to reflect the lower prevalence observed in a specific sub-group like younger adults (Robitaille et al., 2012). We selected error-prone data source prevalence values that were lower than the true population prevalence to reflect evidence from validation studies, which have demonstrated sensitivities less than 1.00 (Coleman et al. 2015; Lix et al., 2006; Quan et al., 2008; Williamson et al., 2014). Data source correlation was chosen to test the effect of moderate and high association between data sources (Frank, 2016). The number of markers for the PSSA method, average correlation amongst markers, and correlation pattern were all chosen to reflect data conditions observed in real-world studies.

Supplementary analyses were conducted for specific combination of simulation conditions, to further explore the performance of our selected methods. Specifically, we considered the case when true prevalence was 20% and outcome prevalence was (18%, 10%). This condition was chosen because it mirrors our numeric example of hypertension case ascertainment. We assessed the effect of low marker prevalence, μ_x, for the PSSA method. Marker prevalence values that ranged between μ_x = 5% and μ_x = 20% were selected (Table A2, Appendix). Secondly, we evaluated the effect of using biased estimates of sensitivity and specificity for the RSSA method. We tested three conditions: (1) when estimates of sensitivity and specificity were equal to the truth, (2) when estimates of sensitivity were 10% below the truth, and (3) when estimates of specificity were 10% below the truth. Finally, we performed additional simulations for each data-combining method to explore the potential confounding of the correlation between data sources and their sensitivity/specificity. We fixed the correlation between the data sources at $\rho_{Y_1 Y_2}$ = 0.70, and considered the case when $prev_{\mathrm{T}}$ = 20%. We then compared the outcome prevalence when $(prev_{Y_1}, prev_{Y_2})$ = (16%, 16%) with sensitivity Sn_{Y_1} = 0.65 and specificity Sp_{Y_1} = 0.96, versus $(prev_{Y_1}, prev_{Y_2})$ = (10%, 10%) with sensitivity Sn_{Y_1} = 0.50 and specificity Sp_{Y_1} = 0.99.

3.3.2 Data generation

The simulation study data were generated with the following process:

1) Using copulas, generate a set of N_x = 8 and N_x = 16 binary markers with average correlations of $\bar{\rho}_x$= 0.00, 0.20 and 0.50, from an exchangeable and unstructured correlation matrix, respectively. Copulas are constructed by specifying the joint distribution of correlated random variables that each follow a standardized uniform distribution, that is, uniformly distributed variables with a minimum of 0 and maximum of 1, denoted as, Uniform(0,1).
2) Generate true disease status from a Bernoulli distribution via a logistic regression model; obtain a true prevalence of $prev_{\mathrm{T}}$= 20% and 10%. Specific values of beta coefficients, β_x, and marker prevalence, μ_x, were used to obtain the true prevalence estimates, as shown in Appendix A (Table A.2). These values were selected based on the odds ratios for previous epidemiological studies to estimate the prevalence of hypertension (Kaplan et al., 2010; Walker et al., 2013).
3) Generate error-prone measures of disease status from step 2 based on pre-selected values of sensitivity (Sn_{Y_j}) and specificity (Sp_{Y_j}), j = 1, 2; obtain prevalence for two error-prone sources ($prev_{Y_1}, prev_{Y_2}$) = (5% to 18%). The mechanism for generating an error-prone disease status measure is a conditional Bernoulli process that can be characterized via the following data generating process (Tennekoon & Rosenman, 2016):

$$Y_1 = \mathrm{P}(D = 1)\ [U < \mathrm{P}(Y_1 = 1|\ D = 1)] + \mathrm{P}(D = 0)\ [U < 1 - \mathrm{P}(Y_1 = 0|\ D = 0)] \qquad (3\text{-}16)$$

where Y_1 is the error-prone measure of disease status, P(D = 1) and P(D = 0) are the indicators of true disease status in the population, P(Y_1= 1| D = 1) and P(Y_1= 0| D = 0) are the sensitivity and specificity of the measure of disease status, respectively, and U is a random variable that follows Uniform(0,1).

4) Calculate disease prevalence using each data-combining method: $prev_{OR}$, $prev_{AND}$, $prev_{RSSA}$ and $prev_{PSSA}$. Repeat this process *K*= 500 times.
5) Calculate the sample mean and variance of the *K* prevalence estimates from each data-combining method (m= OR, AND, RSSA, PSSA) as follows:

$$\overline{prev_{\mathrm{m}}} = \frac{1}{K}\sum_{k=1}^{K} prev_{\mathrm{m}\ (k)} \tag{3-17}$$

$$\sigma^2{}_{prev_{\mathrm{m}}} = \frac{1}{k-1}\sum_{k=1}^{K} \left(prev_{\mathrm{m}\ (k)} - \overline{prev_{\mathrm{m}}}\right)^2 \tag{3-18}$$

where $prev_{\mathrm{m}\ (k)}$ is the prevalence estimated using a data-combining method from the k^{th} replication, and *K* is the total number of replications. *K*=500 was selected to allow for sampling variability.

3.3.3 Simulation performance measures

For each combination of simulation conditions, the data-combining methods were evaluated using the following measures of performance: relative bias (RB) and mean square error (MSE) (Walther & Moore, 2005). RB and MSE values were averaged across the *K*=500 replications. RB was calculated as:

$$\mathrm{RB} = \frac{|\ prev_{\mathrm{T}} - \overline{prev_{\mathrm{m}}}\ |}{prev_{\mathrm{T}}} \text{ x } 100\% \tag{3-19}$$

where $prev_{\mathrm{T}}$ is the true disease prevalence and $\overline{prev_{\mathrm{m}}}$ is the sample mean prevalence for a data-combining method based on 500 replications. MSE was calculated as:

$$\mathrm{MSE}= \sigma^2{}_{prev_{\mathrm{m}}} + |\ prev_{\mathrm{T}} - \overline{prev_{\mathrm{m}}}\ |^2 \tag{3-20}$$

where $\sigma^2{}_{prev_{\mathrm{m}}}$, $prev_{\mathrm{T}}$, $\overline{prev_{\mathrm{m}}}$, are defined in equations (3-18) and (3-19).

To improve readability of small MSE numbers, we multiplied each MSE value by 100. This simulation study was conducted using R programming environment software version R-3.4.4 for Windows (The R Project for Statistical Computing, 2018). The full simulation study program is included in Appendix B.

3.4 Numeric example

3.4.1 Data sources

This study was conducted using linked AHRs and EMRs from fiscal years 2005/2006 to 2008/2009 to ascertain cases of hypertension. A fiscal year extends from April 1 to March 31. The study data sources were from the Manitoba Population Research Data Repository housed at the Manitoba Centre for Health Policy (MCHP). EMRs were linked to AHRs using a unique personal health identification number (PHIN). AHRs and EMRs were used to define case ascertainment algorithms for hypertension, measures of comorbidity, and socio-demographic characteristics of the study cohort.

For this study, AHRs included the population registry, hospital discharge abstracts, physician billing claims, and Drug Program Information Network (DPIN) records. The population registry contains information for all Manitobans registered with the Manitoba Health Services Insurance Plan, including healthcare coverage start and end dates, demographics, and postal codes. Hospital discharge abstracts contain information about discharges from acute and chronic care facilities. Before April 2004, up to 16 diagnosis codes based on the International Classification of Diseases (ICD), Ninth Revision, Clinical Modification (ICD-9-CM) were recorded. The 10th version of the Canadian version of ICD was introduced on April 1, 2004 and captures up to 25 diagnosis codes. Physician billing claims are submitted by fee-for-service physicians to the ministry of health for provider remuneration. Each claim includes the date of service and one three-digit ICD-9-CM code for the diagnosis which best reflects the reason for the visit. The Drug Program Information Network (DPIN) is an electronic, online, point-of-sale database that contains information about prescriptions filled by pharmacies. Each approved drug is assigned a Drug Identification Number (DIN) by the Health Canada Drugs Program unit; DINs can be linked to Anatomical Therapeutic Chemical (ATC) codes which are maintained by the World Health Organization (WHO) Collaborating Centre for Drug Statistics Methodology.

EMRs were from the Manitoba Primary Care Research Network (MaPCReN), a network of family physicians. They include information on health problems, billing data, medications, laboratory results, risk factors, referrals, procedures and socio-demographics (Coleman et al., 2015). MaPCReN is Manitoba's network in the CPCSSN (Williamson et al., 2014).

Chronic disease case ascertainment algorithms were developed and validated for each data source independently. Validated AHR case ascertainment algorithms include the type of data source, diagnostic and prescription drug codes, number of records with the codes, and number of years of data (Lix et al., 2006). Validated case ascertainment algorithms for EMRs include text-based and operational-based definitions. The operational-based definition uses information from selected sections within the EMR including encounter diagnoses, health condition lists, medications, and lab results (Williamson et al., 2014). The health condition list, also known as the problem list, is used by clinicians to record and retrieve a patient's medical history (Singer et al., 2016). The health condition list as well as the encounter diagnoses section contains recorded ICD-9 diagnosis codes. The medication section includes one or more prescriptions medication names with corresponding ATC codes. The lab results section includes all laboratory test results, such as the fasting blood glucose and hemoglobin tests.

3.4.2 Study cohort

The study cohort included Manitoba residents with at least one encounter in EMR data between April 1, 2005 and March 31, 2009 who could be linked to at least one AHR (i.e., hospital records or physician billing claims). To be eligible for inclusion in the cohort, an individual required a minimum of seven years of health insurance coverage before the study index date and seven years of coverage after the index date. Second, an individual was included if he/she was at least 18 years of age as of the index date. The study index date was defined as the date of the first encounter in the EMR data. AHRs are available starting in 1970/71, whereas EMRs are available starting in 1998/99. We chose 2005/06 as the initial study year to allow for a seven-year observation window before the index date for case ascertainment in EMR data. Moreover, 2008/09 was selected as the end year because it allows for seven years of follow-up for case ascertainment in EMR data.

A retrospective cohort for hypertension was constructed for which validated AHR and EMR case ascertainment algorithms were defined. The components of the AHR and EMR case ascertainment algorithms are listed in Table 3.3. To be identified as a hypertension case in AHRs, we determined when individuals satisfied the criteria for AHR case ascertainment for hypertension in the study observation period. Their case date was defined as the earliest point in time at which they met the case ascertainment criteria. We determined when individuals satisfied the criteria for EMR case ascertainment for hypertension over a fourteen-year period because there is no time constraint applied in the EMR case ascertainment algorithm.

Table 3.3: AHR and EMR case ascertainment algorithm for hypertension.

Data source	Contact frequency, source and duration	ICD 9-CM/10-CA diagnosis codes	ATC medication codes
AHR	1+H or 2+P in 2 years	ICD-9-CM: 401-405 ICD-10-CA: I10-I13, I15	
EMR	(2+P in 2 years) or 1+PL or 1+Rx ever	ICD-9-CM: 401-405	C07AB04, C09XA02, C03DB01, C08CA01, C07AB03,C07CB03, C09AA07, C09AA01,C07AG02, C03BA04, C09AA08,C09AA02, C09BA02, C09CA02, C09DA02, C08CA02, C09AA09,C03AA03, C03EA01, C03BA11, C09CA04, C09DA04, C09AA03,C09BA03, C09DA01, C02LB01, C03BA08, C09CA07, C07AA06,C09AA10, C03DB02, C09CA03, C08DA01

Note: AHR= administrative health record, EMR= electronic medical record, H = hospital discharge abstract, P = physician billing claim, PL = problem list, Rx = drug codes; ICD-9-CM/ 10-CA = International Classification of Diseases, 9th Revision, Clinical Modification and 10th version of the Canadian version; ATC = Anatomic, Therapeutic, Chemical.

3.4.3 Study variables

Case ascertainment markers were used as model covariates for the PSSA method. Case ascertainment markers included:

(a) Socio-demographic characteristics: sex, age group (18-44, 45-64, 65+ years), income quintile (Q), and region. These variables were defined as of the study index date. Q is an area-level measure of socioeconomic status defined using Statistics Canada Census data. It is based on total household income for dissemination areas, the smallest geographic unit for which Census data are publicly released (Mustard et al., 1997). Q_1 through Q_5 each represent approximately 20% of the total Manitoba population; separate quintiles are defined for rural and urban residents, but were combined into a single quintile for this study. Postal codes from the population registry were used to assign individuals to income quintiles. Region was based on regional boundaries and was defined as Winnipeg and non-Winnipeg.

(b) Charlson comorbidity score (CCS): CCS is a summary measure of a patient's comorbidity that takes into account a number of comorbid conditions based on ICD diagnosis codes from hospital discharge abstracts (Quan et al., 2005). Each comorbid condition has an associated weight and the sum of all the weights results in a single comorbidity score for a patient. We defined CCS as a categorical variable that took on values of 0 (no comorbid conditions), 1-2 (one to two comorbid conditions) and 3+ (three or more comorbid conditions). CCS was calculated for the one-year period prior to the study index date using diagnoses in both the hospital and physician data.

c) Disease-specific case ascertainment markers: chronic obstructive pulmonary disease (COPD), diabetes, depression, dementia, obesity, cerebrovascular disease (CD), congestive heart failure (CHF), coronary heart disease (CHD), renal disease (RD), and substance abuse (SA). These markers were used as independent variables in a logistic regression model to define hypertension cases in Peng et al. (2005), as risk factors for hypertension risk-prediction models in Echouffo-Tcheugui et al. (2013) and Sun et al. (2017). In this study, the first five diseases from the above list were defined from both AHRs and EMRs, while the last five diseases were defined from AHRs only. This was due to the fact that EMR case ascertainment algorithms for the last five diseases have not been developed. AHR-defined diseases were calculated based on a two-year period prior to the index date. EMR-defined diseases were calculated based on a 14-year period (i.e., seven years before and after the index date) because the case definitions are not based on a specified period of time. We defined each disease as a binary variable with values of 1 representing disease present and 0 representing disease absent. Obesity was defined as a dichotomous variable with values of obese (body mass index > 30.0), not obese (body mass index ≤ 30.0), and missing.

3.4.4 Statistical Analysis

Descriptive analyses of the case ascertainment markers were conducted using frequencies, percentages, and correlations. Tetrachoric and polychoric correlations (Juras & Pasaric, 2006) were calculated to measure the associations amongst the binary and categorical case ascertainment markers.

Cohen's kappa (κ) statistic was used to estimate agreement between AHR and EMR case ascertainment algorithms overall, by sex, and by age group. We calculated κ with 95% confidence intervals The following criteria were used to assess the magnitude of agreement: $\kappa < 0.20$ is poor agreement, $0.20 \leq \kappa \leq 0.39$ is fair agreement, $0.40 \leq \kappa \leq 0.59$ is moderate agreement, $0.60 \leq \kappa \leq 0.79$ is good agreement, and $\kappa \geq 0.80$ is very good agreement (Altman, 1990). As well, tetrachoric correlations were used to estimate the relationship between AHRs and EMRs for the entire cohort, by sex and by age group. Stratification was done because the variable association may not stay constant across sex and age.

For each data-combining method, overall disease prevalence estimates and 95% confidence intervals were calculated. We also calculated sex- and age-stratified disease prevalence estimates and 95% confidence intervals. Model fit was assessed for the PSSA method containing different sets of case ascertainment markers by using penalized measures of the log of the likelihood function, the Deviance Information Criterion (DIC; Spiegelhalter et al., 2002). DIC is based on deviance, which is twice the negative observed data log-likelihood denoted as:

$$D(\theta) = -2 \log \mathrm{P}(y \mid \theta) \qquad (3\text{-}21)$$

where $\mathrm{P}(y \mid \theta)$ is the likelihood function with θ and y being the parameter vector and the observed data, respectively. The DIC is then defined as:

$$\mathrm{DIC} = \overline{D}(\theta) + \ p_D \qquad (3\text{-}22)$$

where $\overline{D}(\theta) = \mathrm{E}\{D(\theta)\}$ is the posterior mean of the deviance, $p_D = \overline{D}(\theta) - D(\bar{\theta})$ is the effective number of model parameters with $D(\bar{\theta})$ being the deviance evaluated at the posterior mean of the parameters.

The two components $\overline{D}(\theta)$ and p_D measure model goodness of fit and complexity, respectively. Smaller values of the DIC indicate a better fitting model. However, if two competing models differ in DIC by less than three units, the models are not considered statistically different (Gelman et al., 2014).

CHAPTER 4 – RESULTS FOR SIMULATION STUDY

This chapter describes the results from the simulation study in three sections. The first two sections present the simulation performance measures and the estimated prevalence for each data-combining method when true prevalence is 20% and 10%, respectively. The third section provides an overall comparison of the data-combining methods.

4.1 True prevalence is 20%

4.1.1 Scenario 1: Number of case ascertainment markers (N_x) is 16

The simulation results are described for each of the measures of RB and MSE. Tables 4.1 and 4.2 present the results when the number of case ascertainment markers was set to N_x = 16 and N_x = 8, respectively.

Table 4.1 reveals that RB ranged from 0.9% to 108.8% and MSE ranged from 0.00 to 5.36 across the simulation conditions that we considered. The 'OR' method resulted in the smallest RB and MSE values for outcome prevalence (18%, 10%). For the 'AND', RSSA and PSSA methods, the RB and MSE values were smallest for outcome prevalence (18%, 15%). The RSSA method had the smallest RB (7.5% and 1.1% on average) when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85 and $\rho_{y_1 y_2}$ = 0.65 and the 'OR' method resulted in RB that were the smallest (0.5% and 10.3% on average).

When the average marker correlation was either $\bar{\rho}_x$ = 0.00 and $\bar{\rho}_x$ = 0.20, the PSSA method had the smallest RB (3.4% on average) when the correlation between the data sources $\rho_{y_1 y_2}$ = 0.85 and outcome prevalence was (15%, 15%). As the average marker correlation increased from $\bar{\rho}_x$ = 0.00 to $\bar{\rho}_x$ = 0.50, the RB and MSE values for the PSSA method increased substantially (by more than 90%) irrespective of the correlation between the data sources.

Moreover, the RB showed very little variation (less than 7%) when the average marker correlation was $\bar{\rho}_x$ = 0.00 compared to $\bar{\rho}_x$ = 0.20.

Except for the RSSA method, all data-combining methods had substantially smaller RB and MSE when $\rho_{y_1y_2}$ = 0.85 compared to when $\rho_{y_1y_2}$ = 0.65, regardless of the outcome prevalence.

Table 4.1: Relative bias (RB) and mean squared error (MSE) when true prevalence is 20% and N_x= 16

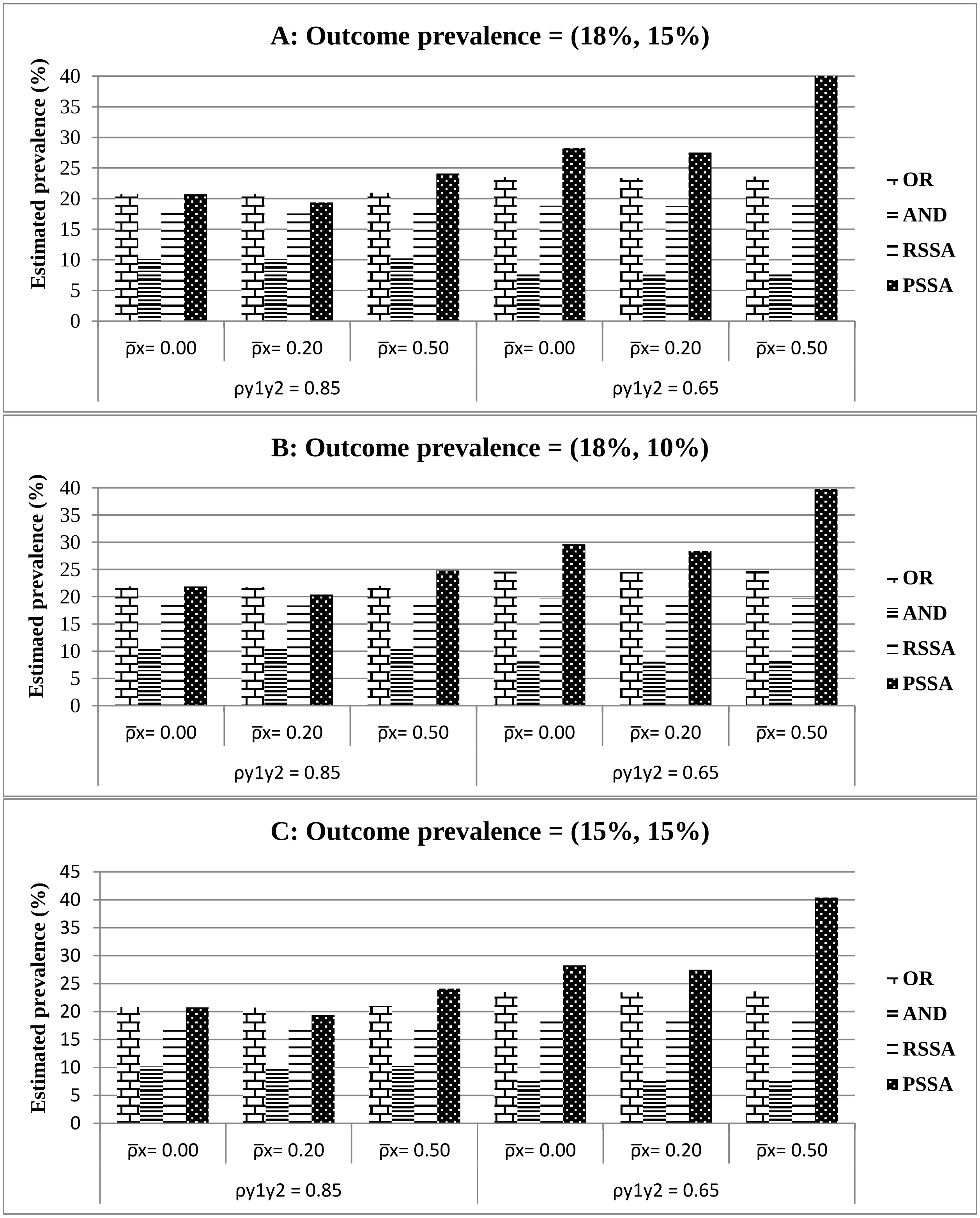
Outcome prevalence ($prev_{Y_1}, prev_{Y_2}$)	$\bar{\boldsymbol{\rho}}_x$	**RB (%)**							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	9.5	47.5	7.5	9.5	23.1	59.4	1.3	48.3
	0.20	9.0	47.7	7.9	2.1	22.9	59.4	1.5	41.7
	0.50	10.1	47.2	7.0	24.3	23.7	59.1	0.9	99.0
18%, 10%	**0.00**	0.3	58.6	18.2	1.1	10.8	67.1	12.8	28.9
	0.20	0.9	58.9	18.7	5.9	10.5	67.2	13.0	31.1
	0.50	0.2	58.3	17.7	48.8	11.2	66.8	12.4	108.8
15%, 15%	**0.00**	4.1	49.2	11.5	3.7	17.6	61.7	5.8	41.3
	0.20	3.6	49.4	12.0	3.1	17.2	61.9	6.1	37.6
	0.50	4.8	48.7	10.9	20.5	18.1	61.5	5.3	102.0
		MSE							
		$\boldsymbol{\rho_{y_1y_2}}$**= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	0.04	0.90	0.02	0.06	0.22	1.41	0.00	0.99
	0.20	0.03	0.91	0.03	0.02	0.21	1.41	0.00	0.82
	0.50	0.04	0.89	0.02	1.06	0.23	1.40	0.00	4.68
18%, 10%	**0.00**	0.00	1.37	0.13	0.02	0.05	1.80	0.07	0.40
	0.20	0.00	1.39	0.14	0.06	0.05	1.80	0.07	0.70
	0.50	0.00	1.36	0.13	2.28	0.05	1.79	0.06	5.36
15%, 15%	**0.00**	0.01	0.97	0.05	0.03	0.13	1.53	0.01	0.74
	0.20	0.01	0.98	0.06	0.02	0.12	1.53	0.02	0.74
	0.50	0.01	0.95	0.05	1.03	0.13	1.51	0.01	4.84

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity adjusted method; PSSA= probabilistic-based sensitivity-specificity adjusted method

Figure 4.1 depicts the estimated prevalence for all data-combining methods when true prevalence is 20% and N_x = 16. Panels A, B and C represent the following combinations of outcome prevalence: (18%, 15%), (18%, 10%) and (15% 15%). Regardless of outcome prevalence, the estimated prevalence for the 'OR' method was the closet to the true prevalence of 20% when $\rho_{y_1 y_2}$ = 0.85; estimated prevalence ranged from 19.9% to 21.9%. However, when $\rho_{y_1 y_2}$ = 0.65, the RSSA method was the best, with estimated prevalence ranging from 18.8% to 19.8%.

When $\bar{\rho}_x$ = 0.00 and $\rho_{y_1 y_2}$ = 0.85, the estimated prevalence for the PSSA method was the closet to the truth (20.8% on average). However, when $\rho_{y_1 y_2}$ = 0.65, the estimated prevalence increased by an average of 35% across the three outcome prevalence conditions. Moreover, when $\bar{\rho}_x$ increased from 0.00 to 0.50, the estimated prevalence increased considerably, especially when $\rho_{y_1 y_2}$ = 0.65. For example, for outcome prevalence combination (18%, 15%), the estimated prevalence increased from 28.3% to 40.4%. The estimated prevalence for the 'AND' method ranged from 10.2% to 10.6% when $\rho_{y_1 y_2}$ = 0.85, and from 7.6% to 8.2% when $\rho_{y_1 y_2}$ = 0.65.

Figure 4.1: Estimated prevalence for data-combining methods when true prevalence is 20% and N_x = 16



4.1.2 Scenario 2: Number of case ascertainment markers (N_x) is 8

Table 4.2 reveals that RB ranged from 0.3% to 90.1% and MSE ranged from 0.00 to 6.16 across the simulation conditions that we considered. The results had a similar pattern to the results obtained when N_x = 16. However, overall the MSE values for the PSSA method were larger than when N_x = 16.

The RSSA method had the smallest RB (5.7% and 0.30% on average) when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85 and $\rho_{y_1 y_2}$ = 0.65, respectively, and the outcome prevalence was (18%, 15%). The 'OR' and 'AND' methods had the best performance in terms of RB and MSE when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85. The 'OR' method resulted in RB estimates that were the smallest (1.8% on average) when outcome prevalence was (18%, 10%) and the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85 and. When the outcome prevalence was (15%, 15%), the PSSA method had the smallest RB (6.8% on average) when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85 and the RSSA method resulted in the smallest RB (4.2% on average) when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.65.

As the average marker correlation increased from $\bar{\rho}_x$ = 0.00 to $\bar{\rho}_x$ = 0.50, the RB and MSE values of the PSSA method increased substantially (by more than 80%) regardless of the correlation between the data sources. Under outcome prevalence (18%, 10%) and when the correlation between the data sources was $\rho_{y_1 y_2}$ = 0.85, the PSSA method had the smallest RB compared to all other conditions, with a value of 3.0%.

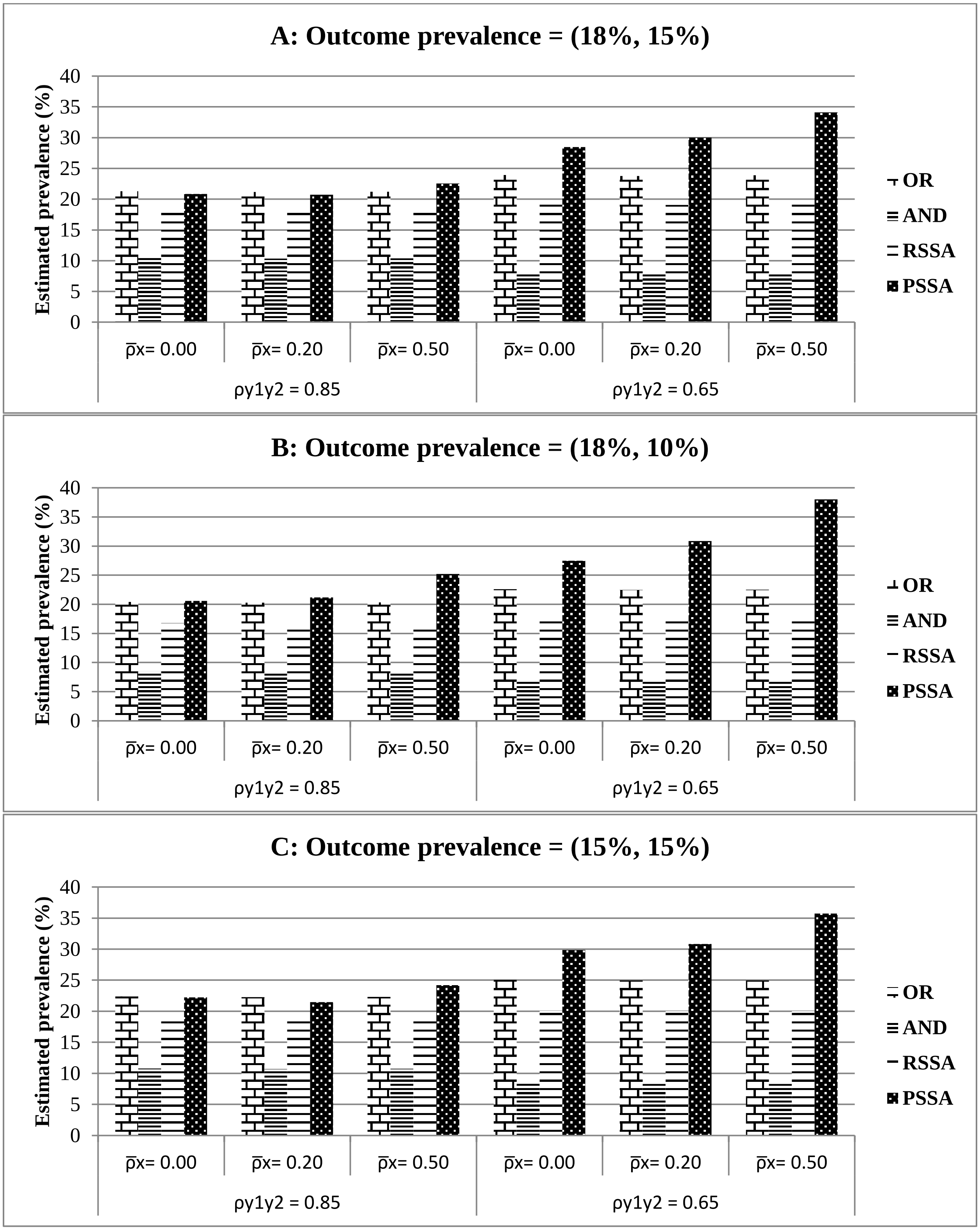
Table 4.2: Relative bias (RB) and mean squared error (MSE) when true prevalence is 20% and N_x= 8

Outcome prevalence ($prev_{Y_1}$, $prev_{Y_2}$)	$\bar{\rho}_x$	**RB (%)**							
		$\rho_{y_1y_2}$ **= 0.85**				$\rho_{y_1y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	12.0	46.0	5.3	11.3	25.2	58.3	0.5	49.5
	0.20	11.4	46.4	5.9	7.4	24.8	58.5	0.1	54.3
	0.50	11.6	46.2	5.7	21.1	25.0	58.4	0.3	78.6
18%, 10%	**0.00**	2.1	57.5	16.2	3.0	13.0	66.1	11.0	37.5
	0.20	1.5	57.8	16.7	5.9	12.3	66.3	11.5	54.3
	0.50	1.7	57.7	16.5	26.1	12.5	66.2	11.4	90.1
15%, 15%	**0.00**	6.5	47.8	9.4	4.3	19.8	60.7	3.9	42.5
	0.20	6.9	48.1	10.0	3.6	18.9	61.0	4.6	50.1
	0.50	7.1	48.0	9.7	12.5	19.5	60.8	4.2	70.7
		MSE							
		$\rho_{y_1y_2}$ **= 0.85**				$\rho_{y_1y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	0.06	0.85	0.01	0.47	0.26	1.36	0.00	1.76
	0.20	0.05	0.86	0.02	0.52	0.25	1.37	0.00	2.25
	0.50	0.06	0.86	0.01	1.47	0.25	1.36	0.00	4.46
18%, 10%	**0.00**	0.00	1.32	0.11	0.69	0.07	1.75	0.05	1.70
	0.20	0.00	1.34	0.11	1.16	0.06	1.76	0.05	3.48
	0.50	0.00	1.33	0.11	2.32	0.06	1.75	0.05	6.16
15%, 15%	**0.00**	0.02	0.91	0.04	0.46	0.16	1.47	0.01	1.62
	0.20	0.02	0.93	0.04	0.72	0.14	1.49	0.01	2.20
	0.50	0.02	0.92	0.04	1.29	0.15	1.48	0.01	3.96

Note: OR= rule-based 'OR' method; AND= rule-based 'AND' method; RSSA= rule-based sensitivity-specificity adjusted method; PSSA= probabilistic-based sensitivity-specificity adjusted method

Figure 4.2 reveals that the prevalence estimates for the 'OR', 'AND' and 'RSSA' methods had a similar trend to that shown in Figure 4.1. The 'OR' method (see Panels A and B), was the closet to the true prevalence of 20% when $\rho_{y_1 y_2}$ = 0.85. The estimated prevalence was on average 21.2% and 20.3% when outcome prevalence was (18%, 15%) and (18%, 10%). However, the estimated prevalence from the RSSA method was the best when outcome prevalence was (15%, 15%) with 18.9%. Moreover, when $\rho_{y_1 y_2}$ = 0.65, the RSSA method was the best with estimated prevalence ranging from 17.7% to 20.1%. There was no difference in the estimated prevalence for the PSSA method when $\rho_{y_1 y_2}$ = 0.85 and $\rho_{y_1 y_2}$ = 0.65 across all outcome prevalence conditions.

Figure 4.2: Estimated prevalence for data-combining methods when true prevalence is 20% and N_x = 8



A: Outcome prevalence = (18%, 15%)

B: Outcome prevalence = (18%, 10%)

C: Outcome prevalence = (15%, 15%)

4.2 True prevalence is 10%

4.2.1 Scenario 1: Number of case ascertainment markers (N_x) is 16

Tables 4.3 and 4.4 present the results when true prevalence is 10% for N_x = 16 and N_x = 8, respectively. Table 4.3 reveals that RB ranged from 0.3% to 333.8% and MSE ranged from 0.00 to 11.79 across the simulation conditions that we considered. The RSSA method had the smallest RB and MSE when outcome prevalence was (8%, 7%), regardless of the correlation between data sources. As outcome prevalence went from (8%, 7%) to (5%, 5%), performance of the RSSA and 'AND' methods got worse. For example, the average RB and MSE for the RSSA method went from 8.2% and 0.01 to 30.8% and 0.1, when the correlation between the data sources was $\rho_{y_1y_2}$ = 0.85. On the other hand, when the correlation between the data sources was $\rho_{y_1y_2}$ = 0.65, the average RB and MSE went from 1.1% and 0.00 to 28.2% and 0.08.

The 'OR' method resulted in RB that were the smallest (0.5% and 10.3% on average) when the outcome prevalence was (8%, 5%) and (5%, 5%) regardless of the correlation between the data sources. For example, for outcome (8%, 5%), the average RB and MSE were 3.3% and 0.00 when the correlation between the data sources was $\rho_{y_1y_2}$ = 0.85, and 12.8% and 0.02, when the correlation between the data sources was $\rho_{y_1y_2}$ = 0.65.

Following similar trends as Tables 4.1 and 4.2, the PSSA method had the smallest RB (1.1% and 6.3%) for outcome prevalence (8%, 5%) and (5%, 5%) when the average marker correlation $\bar{\rho}_x$ = 0.00 and correlation between the data sources was $\rho_{y_1y_2}$ = 0.85. As the average marker correlation increased, the RB and MSE values of the PSSA method increased drastically. For example, under outcome prevalence (8%, 5%) and $\rho_{y_1y_2}$ = 0.85, the PSSA method had RB of 1.1%, 10.7% and 230.5% when the average marker correlation was 0.00, 0.20 and 0.50, respectively.

Table 4.3: Relative bias (RB) and mean squared error (MSE) when true prevalence is 10% and N_x = 16

Outcome prevalence ($prev_{Y_1}$, $prev_{Y_2}$)	$\overline{\rho}_x$	**RB (%)**							
		$\rho_{y_1 y_2}$ **= 0.85**				$\rho_{y_1 y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	11.5	55.8	8.4	9.7	29.6	67.1	1.1	76.2
	0.20	10.5	56.1	9.2	42.6	28.7	67.4	0.3	196.7
	0.50	13.1	54.9	7.1	216.1	30.7	66.6	2.0	307.6
8%, 5%	**0.00**	2.9	59.7	16.0	1.1	12.2	73.4	13.5	45.3
	0.20	3.2	59.7	15.8	10.7	12.0	73.9	13.8	235.2
	0.50	3.8	59.2	15.3	230.5	14.2	73.3	12.1	322.2
5%, 5%	**0.00**	14.7	70.0	30.9	6.3	7.4	78.7	28.4	61.0
	0.20	15.4	70.2	31.5	134.4	8.0	78.7	28.8	271.0
	0.50	13.6	69.5	30.0	275.7	6.3	78.2	27.4	333.8
		MSE							
		$\rho_{y_1 y_2}$ **= 0.85**				$\rho_{y_1 y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	0.01	0.31	0.01	0.28	0.09	0.45	0.00	0.87
	0.20	0.01	0.31	0.01	1.29	0.08	0.45	0.00	5.19
	0.50	0.02	0.30	0.01	5.60	0.10	0.44	0.00	9.97
8%, 5%	**0.00**	0.00	0.36	0.03	0.12	0.02	0.54	0.02	0.28
	0.20	0.00	0.36	0.03	0.59	0.02	0.55	0.02	7.28
	0.50	0.00	0.35	0.02	6.39	0.02	0.54	0.02	10.91
5%, 5%	**0.00**	0.02	0.49	0.10	0.57	0.01	0.62	0.08	1.92
	0.20	0.02	0.49	0.10	4.82	0.01	0.62	0.08	9.11
	0.50	0.02	0.48	0.09	8.71	0.00	0.61	0.08	11.79

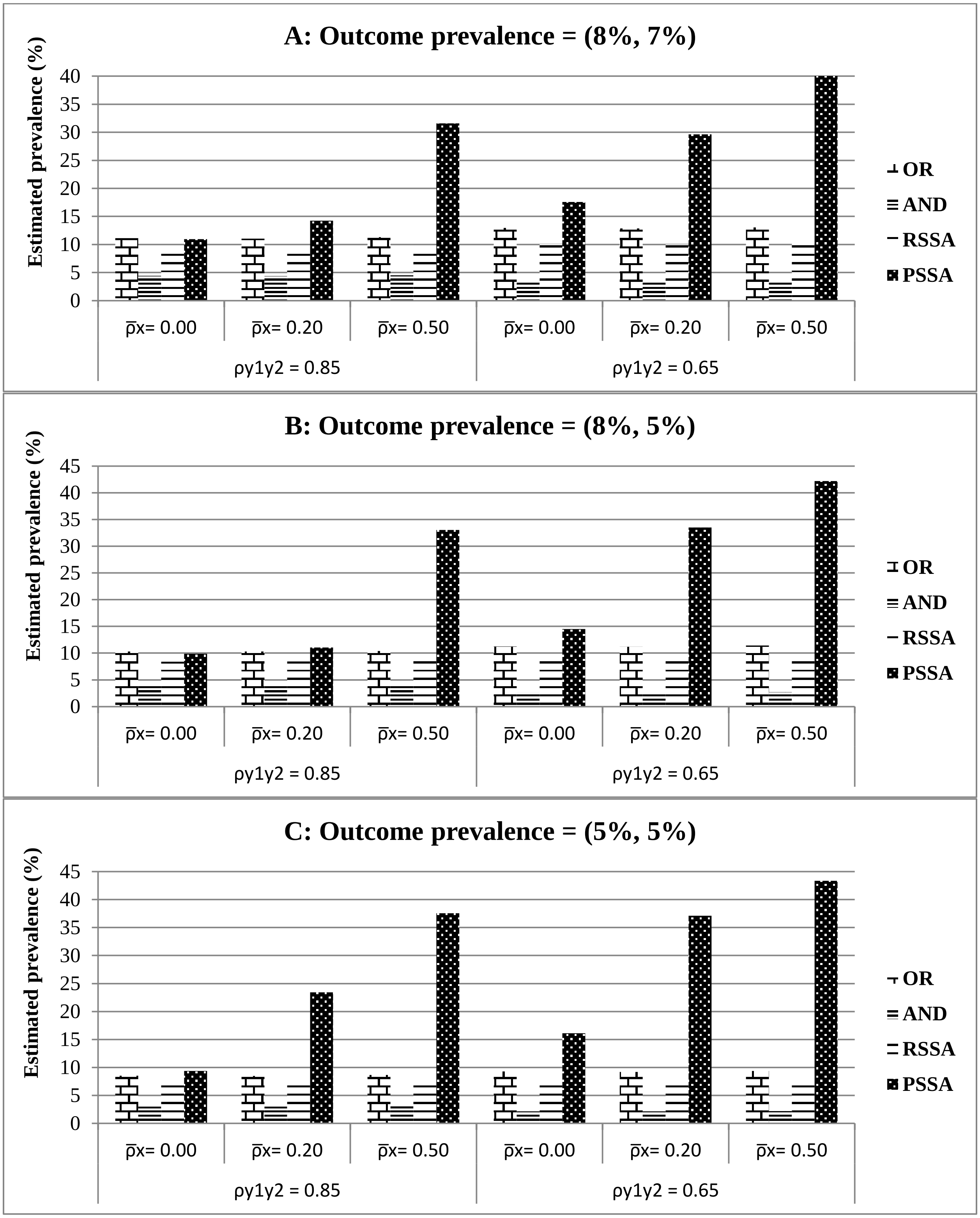
Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity adjusted method; PSSA= probabilistic-based sensitivity-specificity adjusted method

Figure 4.3 displays the estimated prevalence for data-combining methods when true prevalence is 10% and N_x = 16 for three outcome prevalence: (8%, 7%), (8%, 5%) and (5%, 5%). In this condition, the estimated prevalence from the 'OR', 'AND' and 'RSSA' methods showed comparable trend shown in Figures 4.1 & 4.2. However, for the 'OR' method, the estimated was the closet to the true prevalence of 10% for both $\rho_{y_1 y_2}$ = 0.85 and $\rho_{y_1 y_2}$ = 0.65 only when the outcome prevalence was (8%, 5%) and (5%, 5%), with prevalence ranging from 8.5% to 11.4%. On the other hand, the RSSA method was the best at estimated the true prevalence when the outcome prevalence was (8%, 7%) for both $\rho_{y_1 y_2}$ = 0.85 and $\rho_{y_1 y_2}$ = 0, with prevalence ranging from 9.1% to 10.2%.

The estimated prevalence estimated from the 'AND' method was unchanging across the three outcome prevalence conditions, with prevalence ranging from 3.0% to 4.5% when $\rho_{y_1 y_2}$ = 0.85 and 2.1% to 3.3% when $\rho_{y_1 y_2}$ = 0.65.

When the average marker correlation $\bar{\rho}_x$ = 0.00 and $\rho_{y_1 y_2}$ = 0.85, the estimated prevalence from the PSSA method was the closet to the truth (10.1% on average) across all three outcome prevalence conditions. However, when the average marker correlation $\bar{\rho}_x$ increased, the estimated prevalence increased drastically, especially when $\rho_{y_1 y_2}$ = 0.65.

Figure 4.3: Estimated prevalence for data-combining methods when true prevalence is 10% and $N_x = 16$



4.2.2 Scenario 2: Number of case ascertainment markers (N_x) is 8

Table 4.4 reveals that RB ranged from 1.1% to 375.0% and MSE ranged from 0.00 to 18.41 across the simulation conditions that we considered When the number of case ascertainment markers N_x = 8 (Table 4.4), the results showed similar trend to those when the number of case ascertainment markers N_x = 16 (Table 4.3) except for the values of MSE for the PSSA method. In this scenario, the values of MSE for the PSSA method have increased noticeably. For example, the MSE value went from 3.15 to 4.87 when $\rho_{y_1y_2}$ = 0.85 and 6.37 to 10.98 when $\rho_{y_1y_2}$ = 0.65.

Under all of the three outcome prevalence conditions, RB and MSE values of the PSSA method increased as the average marker correlation increased. As the correlation between the data sources went from $\rho_{y_1y_2}$ = 0.85 to $\rho_{y_1y_2}$ = 0.65, the RB and MSE values increased substantially. For example, under outcome prevalence (8%, 7%), the RB and MSE values were 35.0%, 37.8% and 43.3% for average marker correlation of 0.00, 0.20 and 0.50 when $\rho_{y_1y_2}$ = 0.85, and 154.9%, 217.1% and 286.5% when and $\rho_{y_1y_2}$ = 0.65.

The 'OR' method resulted in RB that were the smallest (7.8%, 5.4% and 14.1% on average) across all three outcome prevalence conditions and regardless of the correlation between the data sources except when outcome prevalence was (8%, 7%) and $\rho_{y_1y_2}$ = 0.65. The RSSA method had the smallest RB and MSE (2.0% and 0.00) when outcome prevalence was (8%, 7%) and $\rho_{y_1y_2}$ = 0.65.

As the outcome prevalence went from (8%, 7%) to (5%, 5%), the RSSA and 'AND' methods produced larger values of RB and MSE regardless of the correlation between data sources. For example, the average RB of the 'AND' method were 57.5%, 61.8% and 71.1% for $\rho_{y_1y_2}$ = 0.85, and 68.7%, 74.7% and 79.5% for $\rho_{y_1y_2}$ = 0.65.

Table 4.4: Relative bias (RB) and mean squared error (MSE) when true prevalence is 10% and N_x= 8

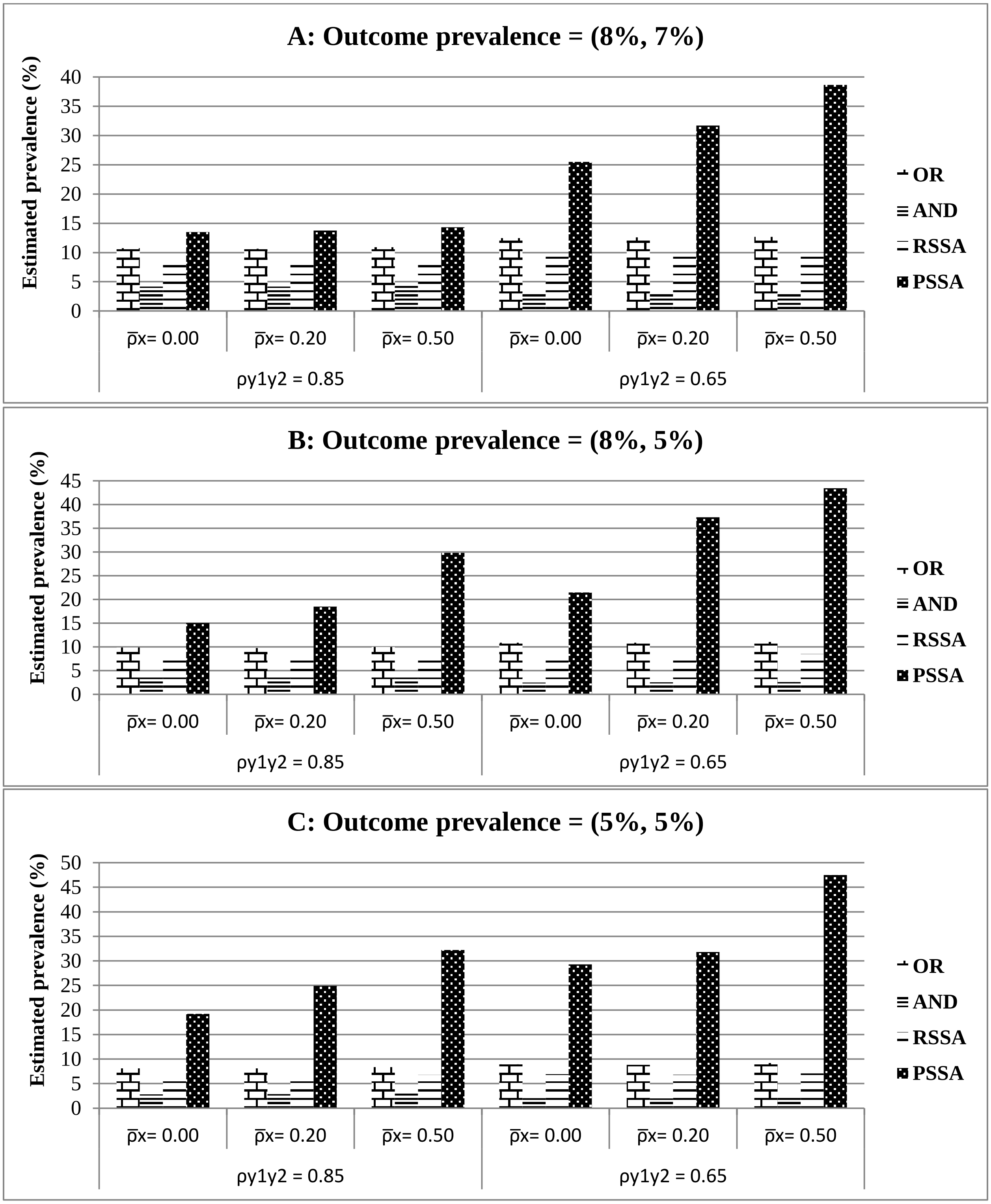
Outcome prevalence ($prev_{Y_1}, prev_{Y_2}$)	$\bar{\rho}_x$	**RB (%)**							
		$\rho_{y_1y_2}$ **= 0.85**				$\rho_{y_1y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	7.5	57.8	11.9	35.0	24.8	69.3	3.0	154.9
	0.20	6.7	58.1	12.5	37.8	25.9	68.6	2.0	217.1
	0.50	9.1	56.6	10.4	43.3	27.1	68.3	1.1	286.5
8%, 5%	**0.00**	1.3	61.5	19.5	50.5	9.1	75.3	16.2	114.9
	0.20	2.5	62.0	20.5	85.1	8.8	74.6	16.2	273.4
	0.50	0.1	62.0	18.8	198.2	10.3	74.2	15.1	334.8
5%, 5%	**0.00**	18.8	71.6	34.3	92.1	10.7	79.4	30.9	193.0
	0.20	18.8	71.3	34.2	149.1	11.9	80.2	32.0	217.9
	0.50	16.4	70.5	32.3	222.1	8.2	78.8	29.0	375.0
		MSE							
		$\rho_{y_1y_2}$ **= 0.85**				$\rho_{y_1y_2}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	0.01	0.33	0.01	2.01	0.06	0.48	0.00	6.33
	0.20	0.01	0.34	0.02	1.31	0.07	0.47	0.00	8.53
	0.50	0.01	0.32	0.01	1.57	0.08	0.47	0.00	12.77
8%, 5%	**0.00**	0.00	0.38	0.04	2.78	0.01	0.57	0.03	4.31
	0.20	0.00	0.38	0.04	3.59	0.01	0.56	0.03	12.63
	0.50	0.00	0.38	0.04	8.45	0.01	0.55	0.02	16.73
5%, 5%	**0.00**	0.04	0.51	0.12	6.96	0.01	0.63	0.10	9.37
	0.20	0.04	0.51	0.12	8.11	0.02	0.64	0.10	9.70
	0.50	0.03	0.50	0.10	10.08	0.01	0.62	0.08	18.41

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity method; PSSA= probabilistic-based sensitivity-specificity method

Figure 4.4 revealed that the prevalence estimates for the ‘OR’, ‘AND’ and ‘RSSA’ methods were comparable to those shown in Figure 4.3. The ‘OR’ method, under Panel B, was the closet to the true prevalence of 10% when the correlation between the data sources $\rho_{y_1 y_2}$ = 0.85, with prevalence ranging from 9.8% to 10.0%. The estimated prevalence from the RSSA method was the best under Panel A with prevalence ranging from 9.7% to 9.9%.

Compared with Figure 4.3, the estimated prevalence from the PSSA method when $\rho_{y_1 y_2}$ = 0.65 increased much faster as the average marker correlation increased in all three outcome prevalence conditions. However, no visible difference was found when $\rho_{y_1 y_2}$ = 0.85.

Figure 4.4: Estimated prevalence for data-combining methods when true prevalence is 10% and N_x = 8



4.3 Comparisons amongst data-combining methods

In regards to the effect of true prevalence, the results have shown an increase in the average RB and MSE for each data-combining method when true prevalence was 10% compared to when it was 20%. Across all simulation conditions, the average RB and MSE were (11.9%, 0.08), (56.7%, 1.31), (8.7%, 0.04), and (35.6%, 1.68) when true prevalence was 20% and (12.7%, 0.63), (68.1%, 3.85), (17.5%, 1.14), and (162.7%, 12.38) when true prevalence was 10%. Based on these results, the RSSA method performed better than other methods when true prevalence was 20% and the 'OR' method performed better than other methods when true prevalence was 10%.

In terms of the effect of the correlation between data sources, the average RB and MSE for the 'OR', 'AND' and PSSA methods became smaller as the correlation increased from $\rho_{y_1 y_2}$ = 0.65 to $\rho_{y_1 y_2}$ = 0.85. The average estimated RB and MSE across all simulation conditions were (17.0%, 0.09), (67.9%, 1.05), (10.5%, 0.03), (141.2%, 5.64) when $\rho_{y_1 y_2}$ = 0.65 and (7.5%, 0.02), (56.9%, 0.73), (15.7%, 0.06), (57.1%, 2.41) when $\rho_{y_1 y_2}$ = 0.85 for the 'OR', 'AND', RSSA and PSSA methods, respectively. The best methods were the RSSA when $\rho_{y_1 y_2}$ = 0.65 and the 'OR' method when $\rho_{y_1 y_2}$ = 0.85.

As for the effect of outcome prevalence, the results revealed that different methods responded differently under the three outcome prevalence conditions. The average estimated RB and MSE across all simulation conditions for the outcome prevalence combination of (18%, 15%) when true prevalence was 20% were (17.4%, 0.14), (52.8%, 1.13), (3.7%, 0.01), (37.3%, 1.55) for the 'OR', 'AND', RSSA and PSSA methods, respectively. Similarly, for the outcome prevalence (18%, 10%) and (15%, 15%), the estimates were (6.4%, 0.03), (62.4%, 1.56), (14.7%, 0.09), (36.8%, 2.03) and (8.7%, 0.05), (58.9%, 1.40), (11.6%, 0.06), (35.8%, 1.63). When true prevalence was 10%, the average estimated RB and MSE for outcome prevalence combination (8%, 7%) was (18.8%, 0.04), (62.2%, 0.39), (5.7%, 0.01), (135.3%, 4.64). Similarly, for outcome prevalence combination (8%, 5%) and (5%, 5%), the estimates were (7.0%, 0.01), (67.4%, 0.46), (16.1%, 0.03), (158.5%, 6.17) and (12.5%, 0.01), (74.8%, 0.51), (30.8%, 0.06), (194.3%, 6.16).

The effect of the average marker correlation on performance of the PSSA method was evident for all simulation conditions. The estimated prevalence became more biased as correlation increased. The average estimated RB and MSE across all simulation conditions were (12.3%, 0.05), (62.5%, 0.89), (13.2%, 0.04), (46.7%, 1.86) when $\bar{\rho}_x$ = 0.00; (12.1%, 0.05), (62.6%, 0.89), (13.5%, 0.05), (90.3%, 3.53) when $\bar{\rho}_x$ = 0.20; and (12.4%, 0.05), (62.1%, 0.88), (12.6%, 0.04), (160.4%, 6.68) when $\bar{\rho}_x$ = 0.50; for the 'OR', 'AND', RSSA and PSSA methods, respectively. The PSSA method performed well when the average marker correlation was low.

In supplementary simulation analyses, we assessed the effect of the marker prevalence on the PSSA method for specific combinations of simulation conditions. We found an increase in the average RB and MSE when the marker prevalence was low. The RB and MSE was (83.2%, 8.95) when marker prevalence ranged between 0.05 and 0.20, versus (40.0%, 2.56) when marker prevalence ranged between 0.10 and 0.80. For the RSSA method, we evaluated the consequence of using biased estimates of sensitivity and specificity on the estimated prevalence. We found that when sensitivity estimates were 10% below the truth, the RB and MSE was (2.7%, 0.04) and when specificity estimates were 10% below the truth, the RB and MSE was (14.1%, 0.08). These results indicate that the RSSA method performs well when estimates of specificity are not underestimated. Finally, we investigated simulation conditions that enabled an exploration of the potential confounding between data source correlation and sensitivity/specificity. When $prev_{\mathrm{T}}$ = 20%, the results (not shown) revealed that all data-combining methods except the PSSA method produced a larger RB and MSE when outcome prevalence was ($prev_{Y_1}$, $prev_{Y_2}$) = (10%, 10%) compared to (16%, 16%). Holding data source correlation at $\rho_{y_1 y_2}$ = 0.70, the average RB increased from 16.0% to 20.5% for the 'OR' method, 56.7% to 74.6% for the 'AND' method, 5.3% to 36.2% for the RSSA method but decreased 66.4% to 48.5% for the PSSA method. These findings suggest that data-combining methods are influenced by the magnitude of the outcome prevalence (i.e., sensitivity and specificity of the data sources).

For all combinations of simulation conditions, similar results were observed when we used an unstructured correlation pattern for the disease markers. Table A.3 to Table A.6 (Appendix A) presents the RB and MSE when true prevalence is 20% and 10%, respectively.

CHAPTER 5 – RESULTS FOR NUMERIC EXAMPLE

Results for the numeric example involving hypertension case ascertainment are described in this chapter. The first and second sections of the chapter present descriptive analyses of the case ascertainment markers in our study cohort. The third section contains model development for the PSSA method. The fourth section presents the prevalence estimates for each of the data-combining methods.

5.1 Description of study cohort

A total of N = 121,144 individuals had at least one encounter in EMRs that could be linked to AHRs (i.e., hospital records or physician billing claims) between April 1, 2005 and March 31, 2009. After exclusions, the study cohort included n = 68,877 individuals.

Cohort members were similar to those in the Manitoba population in terms of sex and age, with a slight over-representation of females and older adults, which are the typical kinds of patients who are most likely to seek health services (Manitoba Government Population Report, 2016). Socio-demographic and CCS characteristics of the study cohort are described in Table 5.1. Close to half of the individuals in the cohort were between 18 and 44 years of age. Slightly more than 55% of the cohort members were female and the majority were Winnipeg residents. Cohort members were equally distributed across the income quintiles, with the exception of the lowest quintiles where they tended to be under-represented. More than 83% of the individuals in the cohort had a CCS of zero.

Disease-specific markers were identified from both AHRs and EMRs. Individuals with depression constituted 10.3% of the study cohort when identified from AHRs and 16.0% when identified from EMRs. A total of 1.9 % of the study cohort had chronic obstructive pulmonary disease when identified from AHRs and 0.3% when identified from EMRs. The fact that 86.0%

of the study cohort is comprised of individuals below 65 years of age may have contributed to the low prevalence of the disease-specific markers. Frequency and percentage of disease-specific case ascertainment markers from AHRs and EMRs are found in Appendix C (Tables C.1 and C.2).

Table 5.1: Socio-demographic and Charlson comorbidity score (CCS) characteristics of the study cohort

Characteristics	**Frequency**	**%**
Sex		
Male	29802	43.3
Female	39075	56.7
Age group		
18-44 years	33007	47.9
45-64 years	26243	38.1
65+ years	9627	14.0
Region		
Non-Winnipeg	30871	44.8
Winnipeg	38006	55.2
Income quintile		
Not found	8888	12.9
Q1 (lowest)	8858	12.9
Q2	10278	14.9
Q3	12154	17.6
Q4	14106	20.5
Q5 (highest)	14593	21.2
CCS		
0	57649	83.7
1 to 2	10348	15.0
3+	880	1.3

Note: Q= Income quintile

5.2 Description of case ascertainment markers

This section describes associations amongst the case ascertainment markers selected for the PSSA model. Measures of agreement and association between AHR and EMR case ascertainment algorithms for hypertension are also described.

There were many case ascertainment markers and the full details of their correlations can be found in Appendix C (see Table C.3). We highlight, here, where there were strong correlations amongst the case ascertainment markers. Let subscripts A and E denote disease-specific markers that were identified from AHRs and EMRs, respectively. The following case ascertainment markers were highly correlated: $\mathrm{CD_A}$ and $\mathrm{COPD_E}$ (ρ_x = -0.99), $\mathrm{Diabetes_A}$ and $\mathrm{Diabetes_E}$ (ρ_x = 0.80), CCS and $\mathrm{Diabetes_A}$ (ρ_x = 0.78), $\mathrm{Dementia_A}$ and $\mathrm{Dementia_E}$ (ρ_x = 0.68), $\mathrm{CHD_A}$ and $\mathrm{CHF_A}$ (ρ_x = 0.61), CCS and $\mathrm{CHF_A}$ (ρ_x = 0.66), and $\mathrm{SA_A}$ and $\mathrm{DM_A}$ (ρ_x = 0.60). We carefully considered selecting combinations of case ascertainment markers for the PSSA model because of high collinearity amongst selected markers. However, overall the mean absolute correlation amongst the markers was low, with $\bar{\rho}_x$= 0.18 and a standard deviation of sd_x = 0.19. When stratifying the study cohort by sex, the mean absolute correlation amongst the markers was $\bar{\rho}_x$= 0.19 for males and $\bar{\rho}_x$= 0.20 for females. When stratifying the study cohort by age group, $\bar{\rho}_x$= 0.21, 0.17 and 0.16 for the 18-44, 45-64 and 65+ age group, respectively.

Table 5.2 displays Cohen's kappa (κ) and the tetrachoric correlation ($\rho_{Y_1Y_2}$) with 95% confidence intervals between AHR and EMR case ascertainment algorithms. Overall there was good agreement with κ = 0.68 (95% CI: 0.67 – 0.68). As well, agreement values for males and females were good, with κ = 0.67 (95% CI: 0.66 – 0.68) for males and κ = 0.68 (95% CI: 0.67 – 0.69) for females. When stratifying by age group, agreement was higher for younger compared to older age groups. The lowest agreement was found for the 65+ age group with κ = 0.45 (95% CI: 0.43 – 0.46).

The association between AHR and EMR case ascertainment algorithms for hypertension was measured using the tetrachoric correlation ($\rho_{Y_1Y_2}$). The correlation was high, with $\rho_{Y_1Y_2}$ = 0.90 (95% CI: 0.89 – 0.90). Similarly, when stratifying the cohort by sex, the degree of association between AHR and EMR case ascertainment algorithms was $\rho_{Y_1Y_2}$ = 0.88 (95% CI: 0.88 – 0.90) for males and $\rho_{Y_1Y_2}$ = 0.90 (95% CI: 0.90 – 0.91) for females. Across age groups, the correlation varied considerably. As age increased the correlation decreased, with $\rho_{Y_1Y_2}$ = 0.89 (95% CI: 0.88 – 0.90) for age group 18-44 years, and $\rho_{Y_1Y_2}$ = 0.76 (95% CI: 0.74 – 0.77) for age group 65+ years.

Table 5.2: Cohen's kappa (κ) and tetrachoric correlations ($\rho_{Y_1Y_2}$) with 95% confidence intervals (CIs) for AHR and EMR case ascertainment algorithms

	$\boldsymbol{\rho_{Y_1Y_2}}$ **(95% CIs)**	$\boldsymbol{\kappa}$ **(95% CIs)**
Overall		
	0.90 (0.89- 0.90)	0.68 (0.67 - 0.68)
Sex		
Males	0.88 (0.88- 0.90)	0.67 (0.66 - 0.68)
Females	0.90 (0.90- 0.91)	0.68 (0.67 - 0.69)
Age group		
18-44 years	0.89 (0.88- 0.90)	0.63 (0.61 - 0.64)
45-64 years	0.87 (0.86- 0.87)	0.64 (0.64 - 0.65)
65+ years	0.76 (0.74- 0.77)	0.45 (0.43 - 0.46)

5.3 Model development for PSSA method

This section describes the results of the model building steps for the PSSA method that was applied to the AHR and EMR data to ascertain hypertension cases. Table 5.3 contains the models and selection criteria. Different subsets of case ascertainment markers were used in model construction based on theoretical evidence and correlation cut-off values for multicollinearity. The maximum number of case ascertainment markers included was 20 (Model 1), and the minimum was 8 (Model 4).

Table 5.3: Models and selection criteria for PSSA method

Model	**Case ascertainment markers**	**Selection criteria**
1	age group, sex, region, Q, CCS, COPDA, COPD_E, DB_A, DB_E, DM_A, DM_E, DP_A, DP_E, OB_A, OB_E, CD_A, CHF_A, CHD_A, RD_A, and SA_A	Include all markers
2	Model 1 excluding COPD_A, DB_A, DM_A, DP_A , OB_A and CCS	Remove redundant markers based on the same measure
3	Model 2 excluding CD_A and CHF_A	Remove markers with $\rho_x > \|0.60\|$
4	age group, sex, COPD_E, DB_E, OB_E, CHF_A, CHD_A, and SA_A	Markers selected based on prior literature

Note: Q= Income quintile; CCS= Charlson Comorbidity Score; DB= Diabetes; CHD= Coronary heart disease; OB= Obesity; COPD= Chronic obstructive pulmonary disease; CD= Cerebrovascular disease; CHF= Congestive heart failure; DP= Depression; DM= Dementia; RD= Renal disease; SA= Substance abuse; Subscripts A and E denote whether the marker was identified from AHRs or EMRs, respectively

For Model 1, all case ascertainment markers were included. For Model 2, COPD, diabetes, dementia, depression and obesity were defined in both AHR and EMR data. Since two measures of the same disease markers were available, only one measure was used. In addition, CCS was calculated based on comorbid conditions, some of which are already identified as disease-specific case ascertainment markers. Therefore, CCS was removed from the modeling.

For Model 3, we excluded markers with very high or very low correlations to avoid potential problems of multicollinearity and model overfitting. For Model 4, the selected markers are those potentially associated with hypertension as suggested by previous literature (Echouffo-Tcheugui et al., 2013; Sun et al., 2017). The selected markers include age, sex, diabetes, obesity, cardiovascular diseases, smoking status and alcohol consumption. COPD was used as a proxy measure for smoking status and substance abuse was used as a proxy measure for alcohol intake.

For each of the PSSA models, visual graphical assessment using trace plots demonstrated that convergence was reached after the 500th iteration; therefore, we run a total of 10,000 iterations of the Gibbs sampler for each of the PSSA models. In addition, we used the Gelman–Rubin diagnostic to ensure the scale reduction, PSRF, of all parameters were smaller than 1.2, suggesting that 10,000 iterations were sufficient for attaining convergence. Once we decided that the chain has converged at iteration 500th, we discarded the first 500 samples as “burn-in” and used the remaining 9,500 samples for inferences. In general, there were no problems with convergence. Figures D.1 to D.6 (Appendix D) display the trace plots and convergence diagnostics for the posterior distribution of our parameter of interest, the disease prevalence.

5.4 Prevalence estimates for data-combining methods

The final two outputs display the estimated hypertension prevalence using each data-combining method for the entire study cohort, by sex and by age group (Tables 5.4 and 5.5). The naïve prevalence of hypertension using AHR and EMR case ascertainment algorithms varied slightly, $prev_{AHRs}$= 30.9% (95% CI: 30.6 – 31.2) versus $prev_{EMRs}$= 24.9% (95% CI: 24.6 – 25.2).

The estimated hypertension prevalence using the OR method was close to the estimate for AHRs alone, $prev_{OR}$ = 34.4% (95% CI: 34.1 – 34.8) versus $prev_{AHRs}$= 30.9% (95% CI: 30.6

– 31.2). This was expected since there was a large amount of overlap between AHRs and EMRs, $\rho_{\boldsymbol{Y_1Y_2}}$ = 0.90 (95% CI: 0.89 – 0.90). Using the RSSA method, the estimated hypertension prevalence was, $prev_{RSSA}$= 32.2% (95% CI: 31.8 – 32.6), which was calculated using values of sensitivity and specificity of the AHRs and EMRs case ascertainment algorithms of hypertension from published Canadian validation studies (Coleman et al. 2015; Kadhim-Saleh et al., 2013; Lix et al., 2006; Quan et al., 2008; Tu et al., 2007; Williamson et al., 2014). Specifically, the sensitivity and specificity values used were: (0.72, 0.95) for AHRs and (0.87, 0.90) for EMRs.

The estimated hypertension prevalence using the PSSA method varied depending on the model applied. The mean absolute correlation amongst the case ascertainment markers included in models 1 through 4 were as follows: $|\bar{\rho}_x|$= 0.18, 0.17, 0.13, and 0.16, respectively. The correlation amongst the case ascertainment markers and prevalence of the markers influenced the prevalence estimates, as shown in our simulation results. PSSA Model 1 produced the highest prevalence estimate compared to all other data-combining methods, $prev_{PSSA_1}$= 35.9% (95% CI: 35.7 – 36.1). This could be due to the high mean absolute correlation amongst the case ascertainment markers in Model 1. On the other hand, PSSA Models 2, 3 and 4 produced prevalence estimates that were lower, with the lowest being Model 4, $prev_{PSSA_4}$= 34.3% (95% CI: 34.1 – 34.5).

When stratifying by sex, the mean absolute correlation amongst the case ascertainment markers included in models 1 through 4 for males were: $|\bar{\rho}_x|$= 0.19, 0.18, 0.14, and 0.19, respectively. Similarly, for females, $|\bar{\rho}_x|$= 0.20, 0.20, 0.16, and 0.20, respectively. Compared to results from the entire cohort, for both males and females, very little difference was found in the terms of the naïve prevalence estimates of hypertension using AHR and EMR case ascertainment algorithms and $prev_{OR}$, $prev_{AND}$, and $prev_{RSSA}$. The lowest estimated prevalence was obtained from PSSA Model 4 for both males and females with, $prev_{PSSA_4}$= 35.1% (95% CI: 34.9 – 35.4) and $prev_{PSSA_4}$= 33.2% (95% CI: 32.9 – 33.5), respectively.

When stratifying by age group, both the naïve prevalence estimates of hypertension and data-combining methods varied across age groups. The naïve prevalence of hypertension using AHR and EMR case ascertainment algorithms were lowest for the 18 – 44 age group, $prev_{AHRs}$= 10.3% (95% CI: 10.0 – 10.6) versus $prev_{EMRs}$= 9.0% (95% CI: 8.7 – 9.3), and highest for the 65+ age group, $prev_{AHRs}$= 75.3% (95% CI: 74.4 – 76.2) versus $prev_{EMRs}$= 56.4% (95% CI: 55.4 – 57.4). Across all age groups, PSSA model 4 produced a lower prevalence estimate compared to PSSA model 1, with the lowest for the 18–44 age group, $prev_{PSSA_4}$= 12.2% (95% CI: 11.9 – 12.4). However, for the 65+ age group, the prevalence estimate was not that low, $prev_{PSSA_4}$= 79.1% (95% CI: 78.8 – 79.5) compared to $prev_{PSSA_1}$= 79.7% (95% CI: 79.4 – 80.0). The degree of correlation between AHRs and EMRs, and the prevalence of the case ascertainment markers may have influenced the prevalence estimates. The mean absolute correlation amongst the case ascertainment markers included in models 1 through 4 for the 18 – 44 age group were: $|\bar{\rho}_x|$= 0.21, 0.20, 0.16, and 0.20, respectively. Similarly, $|\bar{\rho_x}|$= 0.17, 0.16, 0.13 and 0.13 for the 45 – 64 age group, and $|\bar{\rho}_x|$= 0.16, 0.16, 0.13, and 0.17 for the 65+ age group.

In terms of model fit statistics for the PSSA methods, Model 4 resulted in a better fit with the lowest DIC, $\mathrm{DIC}_{\mathrm{PSSA}_4}$ = 165719.00, compared to all other models in the analysis of the entire cohort. For sex stratified analysis, Model 4 had a better fit for both males and females with $\mathrm{DIC}_{\mathrm{PSSA}_4}$ = 72705.69 and $\mathrm{DIC}_{\mathrm{PSSA}_4}$ = 92687.57, respectively. Similarly, for age stratified analysis, the best fitting model was PSSA model 4 (Table 5.6).

In summary, several factors influenced the estimated prevalence of hypertension using each of the data-combining methods. The ‘OR’ and ‘AND’ methods were influenced by the high correlation between AHRs and EMRs of $\rho_{Y_1Y_2}$ > 0.85, except for the 65+ age group. As such, $Prev_{OR}$ produced estimates that were only slightly higher than the naïve prevalence of AHRs, $prev_{AHRs}$, whereas, $prev_{AND}$ produced estimates that were lower than the naïve prevalence of EMRs, $prev_{EMRs}$. As well, each of the PSSA models for the 65+ age group produced prevalence estimates that were high compared to prevalence estimates produced for the other age groups. In addition, the estimated prevalence using PSSA models depended on the types of case ascertainment marker.

Table 5.4: Estimates of hypertension prevalence and 95% confidence intervals (CIs) for data-combining methods, overall and by sex

	Prevalence (95% CIs)		
	Overall	**Males**	**Females**
AHRs	30.9 (30.6 – 31.2)	31.7 (31.2 – 32.2)	30.3 (29.8 – 30.8)
EMRs	24.9 (24.6 – 25.2)	26.0 (25.5 – 26.5)	24.1 (23.7 – 24.5)
OR	34.4 (34.1 – 34.8)	35.7 (35.2 – 36.2)	34.0 (33.5 – 34.5)
AND	21.4 (21.1 – 21.7)	22.1 (21.6 – 22.6)	20.9 (20.5 – 21.3)
RSSA	32.2 (31.8 – 32.6)	33.4 (32.8 – 33.9)	31.3 (30.6 – 31.8)
PSSA model 1	35.9 (35.7 – 36.1)	37.1 (36.8 – 37.3)	34.9 (34.7 – 35.1)
PSSA model 2	35.8 (35.6 – 36.0)	37.0 (36.8 – 37.2)	34.7 (34.5 – 35.0)
PSSA model 3	35.4 (35.3 – 35.6)	36.5 (36.2 – 36.7)	34.5 (34.3 – 34.7)
PSSA model 4	34.3 (34.1 – 34.5)	35.1 (34.9 – 35.4)	33.2 (32.9 – 33.5)

Note: AHRs= Administrative Health Records; EMRs= Electronic Medical Records; OR= rule-based 'OR' method; AND= rule-based 'AND' method; RSSA= rule-based sensitivity-specificity adjusted method; PSSA= probabilistic-based sensitivity-specificity adjusted method

Table 5.5: Estimates of hypertension prevalence and 95% confidence intervals (CIs) for data-combining methods, overall and by age group

	Prevalence (95% CIs)			
	Overall	**18 - 44 years**	**45 - 64 years**	**65+ years**
AHRs	30.9 (30.6 – 31.2)	10.3 (10.0 – 10.6)	40.5 (39.9 – 41.1)	75.3 (74.4 – 76.2)
EMRs	24.9 (24.6 – 25.2)	9.0 (8.7 – 9.3)	33.5 (32.9 – 34.1)	56.4 (55.4 – 57.4)
OR	34.4 (34.1 – 34.8)	12.8 (12.4 – 13.2)	45.3 (44.7 – 45.9)	78.8 (78.0 – 79.6)
AND	21.4 (21.1 – 21.7)	6.4 (6.1 – 6.7)	28.7 (28.1 – 29.3)	53.0 (52.0 – 54.0)
RSSA	32.2 (31.8 – 32.6)	11.9 (11.6 – 12.3)	42.2 (41.6 – 42.8)	73.8 (72.9 – 74.7)
PSSA model 1	35.9 (35.7 – 36.1)	13.9 (13.7 – 14.2)	46.9 (46.7 – 47.3)	79.7 (79.4 – 80.0)
PSSA model 2	35.8 (35.6 – 36.0)	13.6 (13.4 – 13.9)	46.1 (45.9 – 46.4)	79.4 (79.1 – 79.7)
PSSA model 3	35.4 (35.3 – 35.6)	12.8 (12.6 – 13.0)	46.3 (46.0 – 46.6)	79.4 (79.1 – 79.8)
PSSA model 4	34.3 (34.1 – 34.5)	12.2 (11.9 – 12.4)	44.8 (44.5 – 45.1)	79.1 (78.8 – 79.5)

Note: AHRs= Administrative Health Records; EMRs= Electronic Medical Records; OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity adjusted method; PSSA= probabilistic-based sensitivity-specificity adjusted method

Table 5.6: Model fit statistics for the PSSA method, overall, by sex and age group

	DIC					
Model	**Overall**	**Males**	**Females**	**18 - 44 years**	**45-64 years**	**65+ years**
1	167249.40	73417.96	93565.27	42925.35	71311.35	26719.38
2	166994.40	73404.73	93493.06	42920.80	70982.53	26554.06
3	166506.00	73180.58	93350.54	42421.25	71033.17	26621.65
4	165719.00	72705.69	92687.57	42219.99	70483.41	26524.68

Note: DIC= Deviance information criterion

CHAPTER 6 – DISCUSSION AND CONCLUSIONS

6.1 Summary

In this study, four data-combining methods that use information from two error-prone data sources for ascertaining chronic disease cases were compared: (a) rule-based ‘OR’ method, (b) rule-based ‘AND’ method, (c) rule-based sensitivity-specificity adjusted (RSSA) method and (d) probabilistic-based sensitivity-specificity adjusted (PSSA) method. A simulation study was conducted to evaluate the performance of the methods. Then, a numeric example for hypertension case ascertainment was used to demonstrate the methods.

We investigated the following conditions in the simulation study: true population prevalence, error-prone data source prevalence, correlation between data sources, number of markers for PSSA method, average correlation amongst markers and marker correlation pattern. Performance of each data-combining method was assessed using RB and MSE.

Under simulation conditions in which the two data sources were highly correlated, the estimated prevalence from the ‘OR’ method only slightly overestimated the true disease prevalence. On the other hand, for simulation conditions in which the two data sources were not highly correlated, the RSSA method had the lowest RB and MSE among all other data-combining methods. As data source correlation decreased from 0.85 to 0.65, the ‘OR’ method and the PSSA method overestimated the true disease prevalence while the ‘AND’ method underestimated the true prevalence.

The magnitude of true prevalence affected the estimated prevalence for all data-combining methods in the simulation study. The ‘OR’ method performed better than all other data-combing methods when true prevalence was 10% while the RSSA method was best when true prevalence was 20%. In general, as the size of the true prevalence decreased from 20% to

10%, the estimated prevalence for all data combining methods became more biased and less accurate.

The average marker correlation had a significant impact on the bias and accuracy of the prevalence estimates for the PSSA method in the simulation study. When true prevalence was 10%, estimates of RB and MSE increased substantially compared to when true prevalence was 20%. However, number of markers and correlation pattern had little impact on the estimated prevalence. The PSSA method performed better than all other data-combining methods only when correlation between the two data sources was high, the true prevalence was 20% and the average marker correlation was low. Additionally, supplementary analysis showed that the PSSA method performed best when prevalence of the markers was not very low.

In our numeric example, the four data-combining methods were applied to linked AHRs and EMRs to ascertain cases of hypertension. We constructed four PSSA models using different subsets of case ascertainment markers. Case ascertainment markers included socio-demographic characteristics, comorbidity scores, and disease-specific markers defined from AHRs and EMRs. PSSA model 1 included twenty markers selected based on theoretical evidence. Model 2 included a total of sixteen markers after removing selected markers from Model 1 that were potentially redundant. Model 3 included a total of fourteen markers after removing markers with high absolute correlation values. Model 4 included a total of eight markers used often in previous literature to describe/define hypertension.

Our numeric results showed that the estimated prevalence of hypertension using AHR and EMR case ascertainment algorithms were 30.9% and 24.9%, respectively. There was a strong correlation between the two measures, at 0.90. Overall, the estimated prevalence of hypertension from all data-combining methods was higher than the estimates using AHR and

EMR case ascertainment algorithms alone, except for the ‘AND’ method where the prevalence estimate was lower. The estimated prevalence was 21.4% for the ‘AND’ method, 32.2% for the RSSA method, 34.4% for the ‘OR’ method, 35.9% for PSSA model 1, 35.8% for PSSA model 2, 35.4% for PSSA model 3, and 34.3% for PSSA model 4.

When we stratified by sex, the estimated prevalence of hypertension using the data-combining methods were similar to the results obtained for the full cohort. However, when we stratified by age groups, the estimated prevalence using AHR and EMR case ascertainment algorithms varied substantially, with low estimates for the 18-44 age group, 10.3% and 9.0% and very high estimates for the 65+ age group, 75.3% and 56.3%. For the 18-44 age group, the prevalence estimates using data-combining methods ranged from 6.4% for the ‘AND’ method to 13.9% for PSSA model 1. For the 65+ age group, the prevalence estimates ranged 53.0% for the ‘AND’ method to 79.7% for PSSA model 1. Each of the four PSSA models produced prevalence estimates that were consistently high for the 65+ age group compared to the estimates for the younger age groups. The low correlation found between the data sources for the 65+ age group may have resulted in the overestimation of prevalence, as supported by the results of our simulation study.

6.2 Discussion

AHRs and EMRs are the two main data sources used for chronic disease surveillance in Canada. However, these data sources are prone to misclassification errors and the magnitude of error may not be the same in each source. It is difficult to choose a single source to use for chronic disease surveillance, since there are strengths and limitations to each source. Rather than using one data source or the other, a more effective approach might be to combine information from both sources to build on their strengths (He et al., 2014; Reitsma et al., 2009; Zheng et al., 2006). This study is the first to use linked AHRs and EMRs to ascertain individuals with hypertension using data-combining methods. As well, this research revealed, via the simulation study, the strengths and limitations of rule-based and probabilistic-based data-combining methods to assist researchers in selecting an appropriate method based on their data characteristics. Accurate disease case ascertainment is important not only for obtaining accurate prevalence estimates but also for producing unbiased epidemiologic and clinical studies about disease outcomes.

We found a high correlation between the AHR and EMR case ascertainment algorithms for hypertension, which left a limited margin of improvement for the data-combining methods. Other studies have found a high degree of association between data sources for conditions with well-defined diagnostic criteria including hypertension and diabetes (Frank, 2016; Zellweger et al., 2014). In our study cohort, the naïve estimates of hypertension prevalence from AHRs (30.9%) and EMRs (24.9%) were higher than those obtained from other Canadian studies which ranged from 19.6% to 21.3% for AHRs (Pace et al., 2017; Quan et al., 2009; Robiaille et al., 2012) and 22.8% for EMRs (Godwin et al., 2015), but was consistent with one study with 32.0% for AHRs (Tu et al., 2007). However, the patterns in terms of sex and age stratified prevalence estimates were consistent with previous studies (Peng et al., 2015; Robiaille et al., 2012; Tu et al., 2007), which lends face validity to our findings.

Prevalence estimates for the PSSA models were slightly different, but close to the estimate for the 'OR' method. The low variation in prevalence estimates could be attributed to the fact that the mean absolute correlation amongst the markers was moderately low, ranging between 0.13 and 0.18 across the PSSA models. Our simulation study revealed that when the average correlation amongst the marker was zero (i.e., independent markers), the PSSA method produced prevalence estimates that were comparable to when the average correlation amongst the marker was 0.20. Moreover, the average marker correlation was low in our numeric example, meaning that each of the markers was providing unique information to the model.

6.3 Conclusions and recommendations

A number of conclusions and recommendations arise from both the simulation study and the numeric example. Overall, the choice of a data-combining method depends on the characteristics of the data. No single method is preferred. For example, although the 'AND' method is an inherently conservative method, it depends heavily on the data source correlation, and therefore may produce estimates that are not that different from other methods.

With respect to the simulation study, we found that correlation between the two data sources had a substantial impact on estimates of disease prevalence for all data-combining methods. Increasing the correlation reduced the RB and MSE. It is important for researchers to carefully consider the amount of correlation between data sources when attempting to estimate disease prevalence using any of the data-combining methods. When correlation between data sources is very high, using the 'OR' method or the 'AND' method will result in comparable estimates of prevalence. When correlation is low, however, we recommend using the 'OR' method when the sensitivities of the two data sources are low, that is, when both data sources tend to capture true disease cases poorly. And, if both data sources tend to poorly capture true non-disease cases, then the 'AND' method is preferable.

In our simulation study, the RSSA method produced large RB and MSE when we underestimated the specificity of case ascertainment algorithms compared to when true estimates of specificity of case ascertainment algorithms were defined. Therefore, the RSSA method should be used with caution, if accurate estimates of sensitivity and specificity of case ascertainment algorithms are not available. Moreover, the estimated prevalence from the RSSA method was less biased when the size of the true prevalence was 20% compared to 10%. Thus, we recommended using the RSSA when true prevalence is high, as it is less affected by potentially sparse data.

For the PSSA method, reducing the marker prevalence increased the RB and MSE, regardless of other factors that were considered. As such, we recommend including a rich set of markers when implementing the PSSA methods to estimate disease prevalence, especially when true prevalence is low. This is reasonable because the larger the prevalence the more information is provided to construct this model-based approach. When these quantities are low, problems associated with sparse data (i.e., data with few or no subjects at crucial combinations of variable values) may arise. We recommend adopting the PSSA method when correlation between the two data sources is high, the average marker correlation is low and the true prevalence is high.

In the case of conditions with low population prevalence, our simulation study revealed that the performance of the data-combining methods tends to deteriorate. The average RB and MSE estimates increased for each data-combining method when true prevalence was 10% compared to when true prevalence was 20%. The estimates increased by 0.3% for the 'OR' method, 8.8% for the RSSA method, 11.4% for the 'AND' method and 127.1% for the PSSA method. Our findings indicate that calculating prevalence for rare conditions using any data-combining method poses specific challenges (Hampton et al., 2011).

6.4 Strengths and limitations

This study has some limitations. First, the simulation study focused on only a selected number of simulation conditions. Selecting a broader set of conditions might have revealed different strengths and limitations of each data-combining method. At the same time, we investigated a total of 144 combinations of simulation conditions and selected parameter values that reflect scenarios appearing in real-world data (Kaplan et al., 2010; Padwal et al., 2016; Walker et al., 2013). Thus, the simulation study overall provides a thorough assessment of the RB and MSE of each of the data-combining methods.

With respect to the numeric example, in our cohort development, we required seven years of coverage before and after an individual’s first EMR encounter, in order to ensure that we accurately captured EMR cases of hypertension. The EMR case ascertainment algorithms developed by CPCSSN investigators are not specified over a defined period of time. Requiring all cohort members to have long periods of coverage could result in selection bias, because less healthy individuals would have a lower probability of meeting this criteria.

The main strength of this study was the use of both computer simulation and a real numeric example to examine the performance of the data-combining methods. The computer simulation was used to study the performance of the rule-based and probabilistic-based data-combining methods for known population characteristics and the numeric example demonstrated the application of the methods in the real world. We compared methods using two population-based data sources (i.e., AHRs and EMRs) that are available in many jurisdictions in Canada. Moreover, this research investigated multiple sets of case ascertainment markers when applying the PSSA method.

6.5 Future research

The methods used in this study can be extended to combine more than two data sources. For example, future research could investigate including survey data as a third data source. For example, the population-based Canadian Community Health Survey is often used to produce estimates of prevalence for many chronic conditions, including hypertension, which was the focus of the numeric example used in this study (Muggah et al., 2013). Self-report data from health surveys are prone to recall bias, and therefore will produce biased prevalence estimates on their own. Thus, combining this data source with both AHRs and EMRs might be helpful to epidemiologists and public health staff who routinely use only a single source to report disease prevalence estimates.

The PSSA method used in this study only included case ascertainment markers with complete information. However, case ascertainment markers could potentially be characterized by missing data. Further research could extend this method to account for missingness in the markers (Janssen et al., 2010; Rubin, 1987). Extensions can also be considered for error-prone data sources with distributions other than binary (e.g., nominal, ordinal, or continuous).

Finally, all of the data-combining methods can be applied to other real-world numeric examples. They may be most beneficial for diseases for which there is a low to moderate correlation amongst case ascertainment results from error-prone data sources. Examples include arthritis and mental disorders such as anxiety and depression.

REFERENCES

Alonzo, T. A., & Pepe, M. S. (1998). Using a combination of reference tests to assess the accuracy of a new diagnostic test. *Statistics in Medicine*, *18*(22), 2987-3003.

Altman, D. G. (1990). *Practical Statistics for Medical Research*. CRC Press.

Atwood, K. M., Robitaille, C. J., Reimer, K., Dai, S., Johansen, H. L., & Smith, M. J. (2013). Comparison of diagnosed, self-reported, and physically-measured hypertension in Canada. *Canadian Journal of Cardiology*, *29*(5), 606-612.

Barnett, A. G., Koper, N., Dobson, A. J., Schmiegelow, F., & Manseau, M. (2010). Using information criteria to select the correct variance–covariance structure for longitudinal data in ecology. *Methods in Ecology and Evolution*, *1*(1), 15-24.

Bernatsky, S., Joseph, L., Bélisle, P., Boivin, J. F., Rajan, R., Moore, A., & Clarke, A. (2005). Bayesian modelling of imperfect ascertainment methods in cancer studies. *Statistics in Medicine*, *24*(15), 2365-2379.

Bernatsky, S., Lix, L., Hanly, J. G., Hudson, M., Badley, E., Peschken, C., ... & Bélisle, P. (2011). Surveillance of systemic autoimmune rheumatic diseases using administrative data. *Rheumatology International*, *31*(4), 549-554.

Birtwhistle, R., & Williamson, T. (2015). Primary care electronic medical records: a new data source for research in Canada. *Canadian Medical Association Journal*, *187*(4), 239-240.

Burton, P. R. (1994). Helping doctors to draw appropriate inferences from the analysis of medical studies. *Statistics in Medicine*, *13*(17), 1699-1713.

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

Chib, S., & Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, *85*(2), 347-361.

Coleman, N., Halas, G., Peeler, W., Casaclang, N., Williamson, T., & Katz, A. (2015). From patient care to research: a validation study examining the factors contributing to data quality in a primary care electronic medical record database. *BMC Family Practice*, *16*(1), 11.

Cook, S. J., Blas, B., Carroll, R. J., & Sinha, S. (2017). Two wrongs make a right: addressing underreporting in binary data from multiple sources. *Political Analysis*, *25*(2), 223-240.

Couris, C. M., Polazzi, S., Olive, F., Remontet, L., Bossard, N., Gomez, F., & Trombert, B. (2009). Breast cancer incidence using administrative data: correction with sensitivity and specificity. *Journal of Clinical Epidemiology*, *62*(6), 660-666.

Couris, C. M., Colin, C., Rabilloud, M., Schott, A. M., & Ecochard, R. (2002). Method of correction to assess the number of hospitalized incident breast cancer cases based on claims databases. *Journal of Clinical Epidemiology*, *55*(4), 386-391.

Echouffo-Tcheugui, J. B., Batty, G. D., Kivimäki, M., & Kengne, A. P. (2013). Risk models to predict hypertension: a systematic review. *PloS One*, *8*(7), e67370.

Frank, J. (2016). Comparing nationwide prevalences of hypertension and depression based on claims data and survey data: An example from Germany. *Health Policy*, *120*(9), 1061-1069.

Dai, S., Robitaille, C., Bancej, C., & Loukine, L. (2010). Executive summary-report from the Canadian chronic disease surveillance system: hypertension in Canada, 2010. *Chronic Diseases and Injuries in Canada*, *31*(1), 46-47.

Dendukuri, N., & Joseph, L. (2001). Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. *Biometrics*, *57*(1), 158-167.

Garies, S., Birtwhistle, R., Drummond, N., Queenan, J., & Williamson, T. (2017). Data Resource Profile: national electronic medical record data from the canadian primary care Sentinel Surveillance Network (CPCSSN). *International Journal of Epidemiology, 46*(4), 1091-1092f.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997-1016.

Gelman, A., & Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457-472.

Godwin, M., Williamson, T., Khan, S., Kaczorowski, J., Asghari, S., Morkem, R., ... & Birtwhistle, R. (2015). Prevalence and management of hypertension in primary care practices with electronic medical records: a report from the Canadian Primary Care Sentinel Surveillance Network. *CMAJ Open, 3*(1), E76.

Hadgu, A., Dendukuri, N., & Hilden, J. (2005). Evaluation of nucleic acid amplification tests in the absence of a perfect gold-standard test: a review of the statistical and epidemiologic issues. *Epidemiology, 16*(5), 604-612.

Hampton, K. H., Serre, M. L., Gesink, D. C., Pilcher, C. D., & Miller, W. C. (2011). Adjusting for sampling variability in sparse data: geostatistical approaches to disease mapping. *International Journal of Health Geographics*, *10*(1), 54.

He, Y., & Zaslavsky, A. M. (2009). Combining information from cancer registry and medical records data to improve analyses of adjuvant cancer therapies. *Biometrics*, *65*(3), 946-952.

He, Y., Landrum, M. B., & Zaslavsky, A. M. (2014). Combining information from two data sources with misreporting and incompleteness to assess hospice-use among cancer patients: a multiple imputation approach. *Statistics in Medicine*, *33*(21), 3710-3724.

Janssen, K. J., Donders, A. R. T., Harrell, F. E., Vergouwe, Y., Chen, Q., Grobbee, D. E., & Moons, K. G. (2010). Missing covariate data in medical research: to impute is better than to ignore. *Journal of Clinical Epidemiology*, *63*(7), 721-727.

Joseph, L., Gyorkos, T. W., & Coupal, L. (1995). Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, *141*(3), 263-272.

Juras, J., & Pasaric, Z. (2006). Application of tetrachoric and polychoric correlation coefficients to forecast verification. *Geofizika*, *23*(1), 59-82.

Kadhim-Saleh, A., Green, M., Williamson, T., Hunter, D., & Birtwhistle, R. (2013). Validation of the diagnostic algorithms for 5 chronic conditions in the Canadian Primary Care Sentinel Surveillance Network (CPCSSN): a Kingston Practice-based Research Network (PBRN) report. *The Journal of the American Board of Family Medicine*, *26*(2), 159-167.

Kaplan, M. S., Huguet, N., Feeny, D. H., & McFarland, B. H. (2010). Self-reported hypertension prevalence and income among older adults in Canada and the United States. *Social Science & Medicine*, *70*(6), 844-849.

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., Metge, C., McKeen, N., Moore, D., & Bond, R. (2006). Defining and validating chronic diseases: an administrative data approach. *Winnipeg: Manitoba Centre for Health Policy*.

Manitoba Government Population Report. (2016). *2016 annual population report by the Manitoba Health, Healthy Living and Seniors Population*. Retrieved from http://www.gov.mb.ca/health/population/pr2016.pdf

Martin, D. H., Nsuami, M., Schachter, J., Hook, E. W., Ferrero, D., Quinn, T. C., & Gaydos, C. (2004). Use of multiple nucleic acid amplification tests to define the infected-patient “gold standard” in clinical trials of new diagnostic tests for Chlamydia trachomatis infections. *Journal of Clinical Microbiology*, *42*(10), 4749-4758.

Muggah, E., Graves, E., Bennett, C., & Manuel, D. G. (2013). Ascertainment of chronic diseases using population health data: a comparison of health administrative data and patient self-report. *BMC Public Health*, *13*(1), 16.

Mustard, C. A., Derksen, S., Berthelot, J. M., Wolfson, M., & Roos, L. L. (1997). Age-specific education and income gradients in morbidity and mortality in a Canadian province. *Social Science & Medicine*, *45*(3), 383-397.

Naaktgeboren, C. A., Bertens, L. C., van Smeden, M., de Groot, J. A., Moons, K. G., & Reitsma, J. B. (2013). Value of composite reference standards in diagnostic research. *BMJ*, *347*, 1-9.

O’Donnell, S., & Canadian Chronic Disease Surveillance System (CCDSS) Osteoporosis Working Group. (2013). Use of administrative data for national surveillance of osteoporosis and related fractures in Canada: results from a feasibility study. *Archives of Osteoporosis*, *8*(1-2), 143.

Pace, R., Peters, T., Rahme, E., & Dasgupta, K. (2017). Validity of health administrative database definitions for hypertension: a systematic review. *Canadian Journal of Cardiology*, *33*(8), 1052-1059.

Padwal, R. S., Bienek, A., McAlister, F. A., Campbell, N. R., & Outcomes Research Task Force of the Canadian Hypertension Education Program. (2016). Epidemiology of hypertension in Canada: an update. *Canadian Journal of Cardiology*, *32*(5), 687-694.

Pelletier, C., Dai, S., Roberts, K. C., & Bienek, A. (2012). Report summary, Diabetes in Canada: facts and figures from a public health perspective. *Chronic Diseases and Injuries in Canada*, *33*(1), 53-54.

Pepe, M. S. (1992). Inference using surrogate outcome data and a validation sample. *Biometrika*, *79*(2), 355-365.

Quan, H., Smith, M., Bartlett-Esquilant, G., Johansen, H., Tu, K., & Lix, L. (2012). Mining administrative health databases to advance medical science: geographical considerations and untapped potential in Canada. *Canadian Journal of Cardiology*, *28*(2), 152-154.

Quan, H., Li, B., Duncan Saunders, L., Parsons, G. A., Nilsson, C. I., Alibhai, A., & Ghali, W. A. (2008). Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Services Research*, *43*(4), 1424-1441.

Quan, H., Sundararajan, V., Halfon, P., Fong, A., Burnand, B., Luthi, J. C., ... & Ghali, W. A. (2005). Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Medical Care*, 1130-1139.

Reitsma, J. B., Rutjes, A. W., Khan, K. S., Coomarasamy, A., & Bossuyt, P. M. (2009). A review of solutions for diagnostic accuracy studies with an imperfect or missing reference standard. *Journal of Clinical Epidemiology*, *62*(8), 797-806.

Rogan, W. J., & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, *107*(1), 71-76.

Raghunathan, T. E., Xie, D., Schenker, N., Parsons, V. L., Davis, W. W., Dodd, K. W., & Feuer, E. J. (2007). Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening. *Journal of the American Statistical Association*, *102*(478), 474-486.

Robitaille, C., Bancej, C., Dai, S., Tu, K., Rasali, D., Blais, C., & Casey, J. (2013). Surveillance of ischemic heart disease should include physician billing claims: population-based evidence from administrative health data across seven Canadian provinces. *BMC Cardiovascular Disorders*, *13*(1), 88

Robitaille, C., Dai, S., Waters, C., Loukine, L., Bancej, C., Quach, S., ... & Walker, R. (2012). Diagnosed hypertension in Canada: incidence, prevalence and associated mortality. *Canadian Medical Association Journal*, *184*(1), E49-E56.

Rubin DB (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schenker, N., Raghunathan, T. E., & Bondarenko, I. (2010). Improving on analyses of self-reported data in a large-scale health survey by using information from an examination-based survey. *Statistics in Medicine*, *29*(5), 533-545.

Schenker, N., & Raghunathan, T. E. (2007). Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in Medicine*, *26*(8), 1802-1811.

Schiller, I., Smeden, M., Hadgu, A., Libman, M., Reitsma, J. B., & Dendukuri, N. (2016). Bias due to composite reference standards in diagnostic accuracy studies. *Statistics in Medicine*, *35*(9), 1454-1470.

Singer, A., Yakubovich, S., Kroeker, A. L., Dufault, B., Duarte, R., & Katz, A. (2016). Data quality of electronic medical records in Manitoba: do problem lists accurately reflect chronic disease billing diagnoses? *Journal of the American Medical Informatics Association*, *23*(6), 1107-1112.

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian Approaches to Clinical Trials and Health-Care Evaluation* (Vol. 13). Chichester, UK: John Wiley & Sons.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(4), 583-639.

Sun, D., Liu, J., Xiao, L., Liu, Y., Wang, Z., Li, C., ... & Wen, S. (2017). Recent development of risk-prediction models for incident hypertension: An updated systematic review. *PloS One*, *12*(10), e0187240.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*(398), 528-540.

Tennekoon, V., & Rosenman, R. (2016). Systematically misclassified binary dependent variables. *Communications in Statistics-Theory and Methods*, *45*(9), 2538-2555.

Tu, K., Campbell, N. R., Chen, Z. L., Cauch-Dudek, K. J., & McAlister, F. A. (2007). Accuracy of administrative databases in identifying patients with hypertension. *Open Medicine*, *1*(1), e18.

The R Project for Statistical Computing. (2018). The R Project for Statistical Computing from http://www.r-project.org/

Valle, D., Lima, J. M. T., Millar, J., Amratia, P., & Haque, U. (2015). Bias in logistic regression due to imperfect diagnostic test results and practical correction approaches. *Malaria Journal*, *14*, 434.

Walker, R. L., Chen, G., McAlister, F. A., Campbell, N. R., Hemmelgarn, B. R., Dixon, E., ... & Quan, H. (2013). Hospitalization for uncomplicated hypertension: an ambulatory care sensitive condition. *Canadian Journal of Cardiology*, *29*(11), 1462-1469.

Walther, B. A., & Moore, J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*(6), 815-829.

Williamson, T., Green, M. E., Birtwhistle, R., Khan, S., Garies, S., Wong, S. T., ... & Drummond, N. (2014). Validating the 8 CPCSSN case definitions for chronic disease surveillance in a primary care database of electronic health records. *The Annals of Family Medicine*, *12*(4), 367-372.

Yucel, R. M., & Zaslavsky, A. M. (2005). Imputation of binary treatment variables with measurement error in administrative data. *Journal of the American Statistical Association*, *100*(472), 1123-1132.

Zellweger, U., Bopp, M., Holzer, B. M., Djalali, S., & Kaplan, V. (2014). Prevalence of chronic medical conditions in Switzerland: exploring estimates validity by comparing complementary data sources. *BMC Public Health*, *14*(1), 1157.

Zheng, H., Yucel, R., Ayanian, J. Z., & Zaslavsky, A. M. (2006). Profiling providers on use of adjuvant chemotherapy by combining cancer registry and medical record data. *Medical Care*, *44*(1), 1-7.

APPENDIX A: Details of the Computer Simulation

Table A.1: Summary of simulation conditions

Condition #	$\boldsymbol{prev_{\mathrm{T}}}$	$\boldsymbol{N_x}$	$\boldsymbol{\bar{\rho}_x}$	$\boldsymbol{\bar{\rho}_x}$ **(pattern)**	$\boldsymbol{prev_{Y1}}$	$\boldsymbol{prev_{Y2}}$	$\boldsymbol{\rho_{Y_1Y_2}}$	$\boldsymbol{Sn_{Y_1}}$	$\boldsymbol{Sp_{Y_1}}$	$\boldsymbol{Sn_{Y_2}}$	$\boldsymbol{Sp_{Y_2}}$
1	**0.20**	**16**	0.00	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
2	**0.20**	**16**	0.20	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
3	**0.20**	**16**	0.50	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
4	**0.20**	**16**	0.00	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
5	**0.20**	**16**	0.20	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
6	**0.20**	**16**	0.50	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
7	**0.20**	**16**	0.00	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
8	**0.20**	**16**	0.20	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
9	**0.20**	**16**	0.50	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
10	**0.20**	**16**	0.00	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
11	**0.20**	**16**	0.20	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
12	**0.20**	**16**	0.50	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
13	**0.20**	**16**	0.00	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
14	**0.20**	**16**	0.20	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
15	**0.20**	**16**	0.50	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
16	**0.20**	**16**	0.00	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
17	**0.20**	**16**	0.20	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
18	**0.20**	**16**	0.50	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
19	**0.20**	**16**	0.00	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
20	**0.20**	**16**	0.20	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
21	**0.20**	**16**	0.50	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
22	**0.20**	**16**	0.00	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
23	**0.20**	**16**	0.20	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
24	**0.20**	**16**	0.50	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
25	**0.20**	**16**	0.00	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
26	**0.20**	**16**	0.20	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
27	**0.20**	**16**	0.50	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
28	**0.20**	**16**	0.00	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
29	**0.20**	**16**	0.20	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
30	**0.20**	**16**	0.50	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
31	**0.20**	**16**	0.00	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
32	**0.20**	**16**	0.20	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
33	**0.20**	**16**	0.50	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
34	**0.20**	**16**	0.00	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98
35	**0.20**	**16**	0.20	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98
36	**0.20**	**16**	0.50	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98

37	**0.20**	**8**	0.00	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
38	**0.20**	**8**	0.20	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
39	**0.20**	**8**	0.50	ex	0.18	0.15	0.65	0.72	0.96	0.55	0.95
40	**0.20**	**8**	0.00	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
41	**0.20**	**8**	0.20	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
42	**0.20**	**8**	0.50	un	0.18	0.15	0.65	0.72	0.96	0.55	0.95
43	**0.20**	**8**	0.00	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
44	**0.20**	**8**	0.20	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
45	**0.20**	**8**	0.50	ex	0.18	0.15	0.85	0.80	0.98	0.65	0.98
46	**0.20**	**8**	0.00	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
47	**0.20**	**8**	0.20	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
48	**0.20**	**8**	0.50	un	0.18	0.15	0.85	0.80	0.98	0.65	0.98
49	**0.20**	**8**	0.00	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
50	**0.20**	**8**	0.20	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
51	**0.20**	**8**	0.50	ex	0.18	0.10	0.65	0.65	0.94	0.50	0.99
52	**0.20**	**8**	0.00	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
53	**0.20**	**8**	0.20	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
54	**0.20**	**8**	0.50	un	0.18	0.10	0.65	0.65	0.94	0.50	0.99
55	**0.20**	**8**	0.00	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
56	**0.20**	**8**	0.20	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
57	**0.20**	**8**	0.50	ex	0.18	0.10	0.85	0.82	0.99	0.50	0.99
58	**0.20**	**8**	0.00	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
59	**0.20**	**8**	0.20	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
60	**0.20**	**8**	0.50	un	0.18	0.10	0.85	0.82	0.99	0.50	0.99
61	**0.20**	**8**	0.00	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
62	**0.20**	**8**	0.20	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
63	**0.20**	**8**	0.50	ex	0.15	0.15	0.65	0.68	0.97	0.55	0.95
64	**0.20**	**8**	0.00	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
65	**0.20**	**8**	0.20	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
66	**0.20**	**8**	0.50	un	0.15	0.15	0.65	0.68	0.97	0.55	0.95
67	**0.20**	**8**	0.00	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
68	**0.20**	**8**	0.20	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
69	**0.20**	**8**	0.50	ex	0.15	0.15	0.85	0.71	0.98	0.71	0.98
70	**0.20**	**8**	0.00	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98
71	**0.20**	**8**	0.20	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98
72	**0.20**	**8**	0.50	un	0.15	0.15	0.85	0.71	0.98	0.71	0.98
73	**0.10**	**16**	0.00	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
74	**0.10**	**16**	0.20	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
75	**0.10**	**16**	0.50	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
76	**0.10**	**16**	0.00	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98
77	**0.10**	**16**	0.20	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98

78	**0.10**	**16**	0.50	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98
79	**0.10**	**16**	0.00	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
80	**0.10**	**16**	0.20	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
81	**0.10**	**16**	0.50	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
82	**0.10**	**16**	0.00	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99
83	**0.10**	**16**	0.20	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99
84	**0.10**	**16**	0.50	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99
85	**0.10**	**16**	0.00	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
86	**0.10**	**16**	0.20	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
87	**0.10**	**16**	0.50	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
88	**0.10**	**16**	0.00	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
89	**0.10**	**16**	0.20	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
90	**0.10**	**16**	0.50	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
91	**0.10**	**16**	0.00	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
92	**0.10**	**16**	0.20	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
93	**0.10**	**16**	0.50	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
94	**0.10**	**16**	0.00	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
95	**0.10**	**16**	0.20	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
96	**0.10**	**16**	0.50	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
97	**0.10**	**16**	0.00	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
98	**0.10**	**16**	0.20	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
99	**0.10**	**16**	0.50	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
100	**0.10**	**16**	0.00	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
101	**0.10**	**16**	0.20	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
102	**0.10**	**16**	0.50	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
103	**0.10**	**16**	0.00	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
104	**0.10**	**16**	0.20	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
105	**0.10**	**16**	0.50	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
106	**0.10**	**16**	0.00	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99
107	**0.10**	**16**	0.20	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99
108	**0.10**	**16**	0.50	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99
109	**0.10**	**8**	0.00	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
110	**0.10**	**8**	0.20	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
111	**0.10**	**8**	0.50	ex	0.08	0.07	0.65	0.56	0.97	0.54	0.98
112	**0.10**	**8**	0.00	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98
113	**0.10**	**8**	0.20	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98
114	**0.10**	**8**	0.50	un	0.08	0.07	0.65	0.56	0.97	0.54	0.98
115	**0.10**	**8**	0.00	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
116	**0.10**	**8**	0.20	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
117	**0.10**	**8**	0.50	ex	0.08	0.07	0.85	0.70	0.99	0.59	0.99
118	**0.10**	**8**	0.00	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99

119	**0.10**	**8**	0.20	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99
120	**0.10**	**8**	0.50	un	0.08	0.07	0.85	0.70	0.99	0.59	0.99
121	**0.10**	**8**	0.00	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
122	**0.10**	**8**	0.20	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
123	**0.10**	**8**	0.50	ex	0.08	0.05	0.65	0.70	0.99	0.35	0.98
124	**0.10**	**8**	0.00	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
125	**0.10**	**8**	0.20	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
126	**0.10**	**8**	0.50	un	0.08	0.05	0.65	0.70	0.99	0.35	0.98
127	**0.10**	**8**	0.00	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
128	**0.10**	**8**	0.20	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
129	**0.10**	**8**	0.50	ex	0.08	0.05	0.85	0.75	0.99	0.50	0.99
130	**0.10**	**8**	0.00	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
131	**0.10**	**8**	0.20	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
132	**0.10**	**8**	0.50	un	0.08	0.05	0.85	0.75	0.99	0.50	0.99
133	**0.10**	**8**	0.00	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
134	**0.10**	**8**	0.20	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
135	**0.10**	**8**	0.50	ex	0.05	0.05	0.65	0.50	0.99	0.40	0.98
136	**0.10**	**8**	0.00	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
137	**0.10**	**8**	0.20	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
138	**0.10**	**8**	0.50	un	0.05	0.05	0.65	0.50	0.99	0.40	0.98
139	**0.10**	**8**	0.00	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
140	**0.10**	**8**	0.20	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
141	**0.10**	**8**	0.50	ex	0.05	0.05	0.85	0.53	0.99	0.53	0.99
142	**0.10**	**8**	0.00	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99
143	**0.10**	**8**	0.20	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99
144	**0.10**	**8**	0.50	un	0.05	0.05	0.85	0.53	0.99	0.53	0.99

Table A.2: Disease marker prevalence (μ_x) and beta coefficients (β_x **)** based on true disease prevalence ($prev_{\mathrm{T}}$) and number of markers (N_x) used for data generation

$\boldsymbol{prev_{\mathrm{T}}}$	$\boldsymbol{N_x}$	$\mu_{\boldsymbol{x}}$	$\boldsymbol{\beta_x}$
20%	8	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80	-0.35, 1.20, 0.90, 0.50, 0.15, 0.05, -0.20, -0.80, -1.30
	16	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.50, 0.50, 0.50, 0.50, 0.30, 0.30, 0.70, 0.70	-0.35, 1.10, 0.80, 0.40, 0.15, 0.04, -0.20, -0.35, -1.45, 0.04, 0.04, 0.04, 0.04, 0.40, 0.40, -0.35, -0.35
10%	8	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80	-0.35, 1.10, 0.20, 0.50, 0.15, 0.05, -0.50, -1.30, -1.80
	16	0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.50, 0.50, 0.50, 0.50, 0.30, 0.30, 0.70, 0.70	-0.35, 1.10, 0.20, 0.40, 0.15, 0.02, -0.60, -0.50, -1.80, 0.02, 0.02, 0.02, 0.02, 0.50, 0.50, -0.50, -0.50
0.20*	16	0.05, 0.10, 0.10, 0.05, 0.15, 0.10, 0.10, 0.15, 0.05, 0.15, 0.15, 0.20, 0.15, 0.20, 0.10, 0.20	-0.90, 0.10, 0.05, -0.30, -0.35, -0.10, -0.35, 0.05, -0.40, -0.40, -0.60, 0.10, -0.55, 0.05, -0.35, 0.10, -0.40
	8	0.05, 0.10, 0.10, 0.05, 0.15, 0.10, 0.10, 0.15	-1.00, 0.05, -0.70, -0.30, 0.15, -0.50, -0.60, -0.40, -0.80

Note: * was used for specific combination of simulation conditions

Table A.3: Relative bias (RB) and mean squared error (MSE) when true prevalence is 20% and N_x= 16

Outcome prevalence $\boldsymbol{prev_{y_1}, prev_{y_2}}$	$\boldsymbol{\bar{\rho}_x}$	**RB (%)**							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	9.9	47.1	7.1	7.7	23.9	59.0	0.7	47.9
	0.20	8.1	48.6	8.8	2.4	21.8	59.9	2.4	39.6
	0.50	10.9	46.9	6.4	25.2	24.6	58.9	0.2	101.5
18%, 10%	**0.00**	0.3	58.4	17.7	0.2	11.8	66.5	12.0	27.6
	0.20	1.8	59.4	19.4	7.8	9.3	67.8	14.1	34.4
	0.50	1.3	57.8	16.8	35.7	12.2	66.7	11.8	104.7
15%, 15%	**0.00**	5.0	48.7	10.7	3.8	18.2	61.5	5.3	40.8
	0.20	2.8	49.9	12.7	3.6	16.3	62.3	6.9	36.7
	0.50	5.6	48.3	10.2	26.0	19.1	61.1	4.6	104.4
		MSE							
		$\boldsymbol{\rho_{y_1y_2}}$**= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	0.04	0.89	0.02	0.05	0.23	1.39	0.00	0.97
	0.20	0.03	0.94	0.03	0.01	0.19	1.44	0.00	0.72
	0.50	0.05	0.88	0.02	1.04	0.24	1.39	0.00	4.75
18%, 10%	**0.00**	0.00	1.36	0.13	0.02	0.06	1.77	0.06	0.37
	0.20	0.00	1.41	0.15	0.04	0.04	1.84	0.08	1.26
	0.50	0.00	1.34	0.11	1.57	0.06	1.78	0.06	5.02
15%, 15%	**0.00**	0.01	0.95	0.05	0.03	0.13	1.51	0.01	0.73
	0.20	0.00	1.00	0.07	0.03	0.11	1.55	0.02	0.72
	0.50	0.01	0.93	0.04	1.28	0.15	1.50	0.01	4.87

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity method; PSSA= probabilistic-based sensitivity-specificity method

Table A.4: Relative bias (RB) and mean squared error (MSE) when true prevalence is 20% and N_x= 8

Outcome prevalence $\boldsymbol{prev_{y_1}, prev_{y_2}}$	$\boldsymbol{\overline{\rho}_x}$	**RB (%)**							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	12.8	45.5	4.6	14.9	26.3	58.4	1.2	62.7
	0.20	10.9	46.8	6.4	10.2	24.6	58.8	0.1	79.9
	0.50	11.8	46.0	5.5	37.8	24.5	58.6	0.1	81.4
18%, 10%	**0.00**	3.3	57.1	15.2	14.9	12.9	66.0	11.1	28.7
	0.20	1.4	57.8	16.8	2.1	12.1	66.3	11.7	57.1
	0.50	0.9	57.9	17.1	81.5	12.2	66.2	11.6	85.7
15%, 15%	**0.00**	7.6	47.7	8.6	9.4	20.2	60.4	3.6	48.9
	0.20	5.0	48.6	10.7	1.1	18.8	61.2	4.7	47.5
	0.50	5.3	48.4	10.5	12.6	18.8	61.1	4.8	74.1
		MSE							
		$\boldsymbol{\rho_{y_1y_2}}$**= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
18%, 15%	**0.00**	0.07	0.83	0.01	0.89	0.28	1.37	0.00	3.57
	0.20	0.05	0.88	0.02	1.33	0.24	1.38	0.00	6.13
	0.50	0.06	0.85	0.01	3.19	0.24	1.37	0.00	4.76
18%, 10%	**0.00**	0.01	1.31	0.09	1.85	0.07	1.74	0.05	0.47
	0.20	0.00	1.34	0.11	0.56	0.06	1.76	0.06	3.13
	0.50	0.00	1.34	0.12	6.52	0.06	1.75	0.05	4.76
15%, 15%	**0.00**	0.02	0.91	0.03	0.53	0.16	1.46	0.01	1.89
	0.20	0.01	0.95	0.05	0.27	0.14	1.50	0.01	1.70
	0.50	0.01	0.94	0.04	2.74	0.14	1.50	0.01	4.12

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity method; PSSA= probabilistic-based sensitivity-specificity method

Table A.5: Relative bias (RB) and mean squared error (MSE) when true prevalence is 10% and N_x= 16

Outcome prevalence $\boldsymbol{prev_{y_1}, prev_{y_2}}$	$\boldsymbol{\overline{\rho}_x}$	**RB (%)**							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	11.9	55.5	8.1	12.7	30.0	67.1	1.2	104.4
	0.20	8.7	56.8	10.7	31.3	25.7	68.2	2.1	208.9
	0.50	15.2	54.0	5.3	237.1	32.0	66.2	2.9	326.2
8%, 5%	**0.00**	2.8	59.9	16.2	2.6	12.6	73.2	13.2	70.7
	0.20	0.2	61.2	18.7	18.9	11.1	74.1	14.4	244.1
	0.50	6.8	58.3	13.0	238.9	15.5	73.1	11.2	331.8
5%, 5%	**0.00**	14.9	69.4	30.9	5.0	7.4	78.9	28.4	52.7
	0.20	16.6	70.6	32.5	158.4	9.3	79.4	29.9	290.0
	0.50	11.4	69.2	28.4	299.2	3.1	77.5	25.0	351.2
		MSE							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	0.01	0.31	0.01	0.06	0.09	0.45	0.00	1.93
	0.20	0.01	0.32	0.01	0.88	0.07	0.47	0.00	5.98
	0.50	0.02	0.29	0.00	6.96	0.10	0.44	0.00	10.91
8%, 5%	**0.00**	0.00	0.36	0.03	0.01	0.02	0.54	0.02	1.37
	0.20	0.00	0.38	0.04	0.65	0.01	0.55	0.02	7.98
	0.50	0.01	0.34	0.02	6.96	0.02	0.53	0.01	11.38
5%, 5%	**0.00**	0.02	0.48	0.10	0.88	0.01	0.62	0.08	1.49
	0.20	0.03	0.50	0.11	5.32	0.01	0.63	0.09	10.56
	0.50	0.01	0.48	0.08	9.82	0.00	0.60	0.06	12.70

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity method; PSSA= probabilistic-based sensitivity-specificity method

Table A.6: Relative bias (RB) and mean squared error (MSE) when true prevalence is 10% and N_x= 8

Outcome prevalence $\boldsymbol{prev_{y_1}, prev_{y_2}}$	$\overline{\boldsymbol{\rho}}_{\boldsymbol{x}}$	**RB (%)**							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	8.2	57.6	11.3	51.0	25.9	68.8	2.1	127.6
	0.20	7.2	57.9	12.2	111.1	25.7	68.6	2.2	266.3
	0.50	9.6	57.1	10.2	149.5	27.0	68.7	1.3	315.6
8%, 5%	**0.00**	1.7	61.5	19.8	119.4	8.9	74.9	16.3	112.8
	0.20	2.8	62.3	20.8	78.2	9.3	75.1	16.0	216.8
	0.50	0.1	60.8	18.5	209.0	10.7	74.1	14.7	231.4
5%, 5%	**0.00**	18.8	71.3	34.2	79.9	10.8	80.3	31.2	209.3
	0.20	19.0	71.6	34.4	134.4	11.1	80.1	31.4	285.3
	0.50	16.8	71.1	32.8	252.2	9.3	79.3	29.9	298.0
		MSE							
		$\boldsymbol{\rho_{y_1y_2}}$ **= 0.85**				$\boldsymbol{\rho_{y_1y_2}}$ **= 0.65**			
		OR	**AND**	**RSSA**	**PSSA**	**OR**	**AND**	**RSSA**	**PSSA**
8%, 7%	**0.00**	0.01	0.33	0.01	2.54	0.07	0.47	0.00	3.37
	0.20	0.01	0.34	0.02	4.77	0.07	0.47	0.00	11.65
	0.50	0.01	0.33	0.01	6.57	0.07	0.47	0.00	12.99
8%, 5%	**0.00**	0.00	0.38	0.04	7.46	0.01	0.56	0.03	3.00
	0.20	0.00	0.39	0.04	5.53	0.01	0.56	0.03	7.06
	0.50	0.00	0.37	0.04	9.22	0.01	0.55	0.02	9.13
5%, 5%	**0.00**	0.04	0.51	0.12	5.46	0.01	0.64	0.10	10.13
	0.20	0.04	0.51	0.12	8.75	0.01	0.64	0.10	13.17
	0.50	0.03	0.51	0.11	12.27	0.01	0.63	0.09	11.37

Note: OR= rule-based ‘OR’ method; AND= rule-based ‘AND’ method; RSSA= rule-based sensitivity-specificity method; PSSA= probabilistic-based sensitivity-specificity method

APPENDIX B: Simulation Study Code

The R simulation program that created the simulation data and the rule-based and probabilistic-based models are provided below.

```
********************************************************************
Improving Accuracy of Disease Prevalence Estimates by Combining Information from
Administrative Health Records and Electronic Medical Records
Programmer: Saeed Al-Azazi
August 10, 2018
If you have further requests, you can send an email to alazazis@myumanitoba.ca.
********************************************************************
```

```
# Create an R function incorporating the simulation that and the OR, AND, RSSA and PSSA
data-combining methods
# Load R packages
library(psych)
library(copula)
library(MASS)
library(bayesm)
library(writexl)

thesis = function(N, Sn1, Sp1, Sn2, Sp2, n_xi, xi_corr=c(), xi_dist, xi_mean=c(), xi_betas=c()){
 #     where N= sample size
 #     Sn1, Sp1= sensitivity and specificity for data source 1 (AHRs), respectively.
 #     Sn2, Sp2= sensitivity and specificity for data source 2 (EMRs), respectively.
 #     n_xi = number of markers
 #     xi_corr= markers correlation matrix
 #     xi_dist= markers correlation distribution
 #     xi_mean= markers prevalence
 #     xi_betas= markers beta coefficient values
 # Disease prevalence is dependent on markers: A logistic regression is used
 logistic = function (beta0, beta1, beta2, beta3, beta4, beta5, beta6, beta7, beta8, beta9, beta10,
beta11, beta12, beta13, beta14, beta15, beta16, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12,
x13, x14, x15, x16)
   exp(beta0 + beta1*x1 + beta2*x2 + beta3*x3 + beta4*x4 + beta5*x5 + beta6*x6 + beta7*x7 +
beta8*x8 + beta9*x9 + beta10*x10 + beta11*x11 + beta12*x12 + beta13*x13 + beta14*x14 +
beta15*x15 + beta16*x16)/
   (1 + exp(beta0 + beta1*x1 + beta2*x2 + beta3*x3 + beta4*x4 + beta5*x5 + beta6*x6 +
beta7*x7 + beta8*x8 + beta9*x9 + beta10*x10 + beta11*x11 + beta12*x12 + beta13*x13 +
beta14*x14 + beta15*x15 + beta16*x16))
 beta0= xi_betas[1];    beta1= xi_betas[2]
 beta2= xi_betas[3];    beta3= xi_betas[4]
 beta4= xi_betas[5];    beta5= xi_betas[6]
 beta6= xi_betas[7];    beta7= xi_betas[8]
 beta8= xi_betas[9];    beta9= xi_betas[10]
 beta10= xi_betas[11];  beta11= xi_betas[12]
 beta12= xi_betas[13];  beta13= xi_betas[14];
 beta14= xi_betas[15];  beta15= xi_betas[16];
 beta16= xi_betas[17];

 # Generate correlated binary markers using Copula
 ran<- rCopula(N, normalCopula(xi_corr, dim= n_xi, dispstr= xi_dist))

 # Convert Copula's uniform variables to binary variables
 x1<- qbinom(ran[,1],  1, prob= xi_mean[1])
 x2<- qbinom(ran[,2],  1, prob= xi_mean[2])
 x3<- qbinom(ran[,3],  1, prob= xi_mean[3])
 x4<- qbinom(ran[,4],  1, prob= xi_mean[4])
 x5<- qbinom(ran[,5],  1, prob= xi_mean[5])
 x6<- qbinom(ran[,6],  1, prob= xi_mean[6])
 x7<- qbinom(ran[,7],  1, prob= xi_mean[7])
 x8<- qbinom(ran[,8],  1, prob= xi_mean[8])
 x9<- qbinom(ran[,9],  1, prob= xi_mean[9])
 x10<-qbinom(ran[,10], 1, prob= xi_mean[10])
 x11<-qbinom(ran[,11], 1, prob= xi_mean[11])
 x12<-qbinom(ran[,12], 1, prob= xi_mean[12])
 x13<-qbinom(ran[,13], 1, prob= xi_mean[13])
 x14<-qbinom(ran[,14], 1, prob= xi_mean[14])
 x15<-qbinom(ran[,15], 1, prob= xi_mean[15])
 x16<-qbinom(ran[,16], 1, prob= xi_mean[16])

 # Fit a logistic regression: We except a prevalence estimate of 0.20 or 0.10 in the simulated
population
 prob= logistic(beta0, beta1, beta2, beta3, beta4, beta5, beta6, beta7, beta8, beta9, beta10,
beta11, beta12, beta13, beta14, beta15, beta16,
              x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16)
 True_D= rbinom(N, 1, prob)
 P_True= sum(True_D)/ N; P_True

 # Generate two imperfect measures of true disease status using different values of sensitivity
(Sn) and specificity (Sp).
 AHR= (True_D==1)* (runif(N) < Sn1) + (True_D==0)* (runif(N) < 1- Sp1)
 EMR= (True_D==1)* (runif(N) < Sn2) + (True_D==0)* (runif(N) < 1- Sp2)

 # Two-by-Two table: AHRs x EMRs
 nn <- table(EMR, AHR)[c(2,1),c(2,1)]
 n1<- nn[1]; n2<- nn[2]; n3<- nn[3]

 # Calculate the Tetrachoric correlation
 T_correlation<- as.numeric(tetrachoric(nn)[1])
 P_correlation<- as.numeric(phi(nn))

 # Calculate disease prevalence estimate
 P_AHR = (n1 + n2)/ N
 P_EMR = (n1 + n3)/ N

# Calculate the prevalence using OR and AND methods
 P_OR  <-  (n1 + n2 + n3)/ N
 P_AND <-  (n1)/ N
 # Calculate the prevalence using RSSA method
 #Step1: Calculate weights using known Sn and Sp values. From published literature about
hypertension, Sn and Sp for AHRs:(0.72, 0.95); for EMRs:(0.87, 0.90).
 Sn1.r<- 0.72; Sp1.r<- 0.95; Sn2.r<- 0.87; Sp2.r<- 0.90
 p1_01= (1 - Sn1.r)* Sn2.r
 p1_10= Sn1.r * (1 - Sn2.r)
 p0_01= Sp1.r * (1 - Sp2.r)
 p0_10= (1 - Sp1.r) * Sp2.r
 Discor= (n2 + n3)

 #Step2: Make adjustments for the frequencies in the discordant cells
 Weight_D.1= p1_10 * n2+ p1_01 * n3
 Weight_D.0= p0_10 * n2+ p0_01 * n3
 Weight_ratio_D.1= Weight_D.1/ (Weight_D.1 + Weight_D.0)
 sum_D.11= Discor * Weight_ratio_D.1
 P_RSSA=  (sum_D.11 + n1)/ N

# PSSA: Probabilistic-based sensitivity-specificity adjusted method
# First, prepare dataset for analysis
 Dataset_f= as.matrix(cbind(AHR, EMR, x1, x2, x3, x4, x5, x6, x7, x8, x9, x10, x11, x12, x13,
x14, x15, x16))  #Note: To test 8 markers, deleted markers x9 to x16

 # Ready for analysis
 n_total= N              # number of cases in the dataset.
 n_col= 16 + 2           # number of variables in the dataset.
 p_cov= 16 + 1;          # number of covariates for outcome model including the intercept.
 q_1_cov= 16 + 1;        # number of covariates for AHRs model including the intercept.
 q_2_cov= 16 + 1;        # number of covariates for EMRs model including the intercept.

 # Extract the covariate matrix including the intercept:
 x_1_out= cbind(rep(1,n_total), Dataset_f[,3:(p_cov-1+2)]); # For the outcome model
 x_1_rep= x_1_out;      # For the AHRs model
 x_2_rep= x_1_out;      # For the EMRs model
```

```
  # Extract the number of each AHRs/EMRs combination, 11 (Yes Yes), 01 (No Yes), 10 (Yes No),
00 (No No).
  n_total_00= sum(AHR==0 & EMR==0);
  n_total_01= sum(AHR==0 & EMR==1);
  n_total_10= sum(AHR==1 & EMR==0);
  n_total_11= sum(AHR==1 & EMR==1);

  # Extract subsets of the covariates according to each AHRs/ EMRs combination. This step is
needed to calculate the conditional probabilities.
  x_1_out_00= x_1_out[(AHR==0  & EMR==0),];
  x_1_out_01= x_1_out[(AHR==0  & EMR==1),];
  x_1_out_10= x_1_out[(AHR==1  & EMR==0),];
  x_1_out_11= x_1_out[(AHR==1  & EMR==1),];

  x_1_rep_00= x_1_rep[(AHR==0  & EMR==0),];
  x_1_rep_01= x_1_rep[(AHR==0  & EMR==1),];
  x_1_rep_10= x_1_rep[(AHR==1  & EMR==0),];
  x_1_rep_11= x_1_rep[(AHR==1  & EMR==1),];

  x_2_rep_00= x_2_rep[(AHR==0  & EMR==0),];
  x_2_rep_01= x_2_rep[(AHR==0  & EMR==1),];
  x_2_rep_10= x_2_rep[(AHR==1  & EMR==0),];
  x_2_rep_11= x_2_rep[(AHR==1  & EMR==1),];

# Set up initial values for the beta parameters. Use flat prior, which is a prior distribution that
assigns equal likelihood on all possible values of the parameter; the prior could be set to
uniform with some common sense boundaries.
  beta_1_out=  2.00 * runif(p_cov);
  beta_1_rep=  2.00 * runif(q_1_cov);
  beta_11_rep= 2.00 * runif(q_1_cov);
  beta_2_rep=  2.00 * runif(q_2_cov);
  beta_22_rep= 2.00 * runif(q_2_cov);

# Initialize the probabilities in the first step of the Data augmentation (DA) algorithm.
 # Calculate posterior by updating our prior belief with the information from given data.
```

```
  # p_ystar_00: probability of ystar is 1 given both y1 and y2 are zero P(Y*=1|Y1=0, Y2=0, X).
 p_zd_00= pnorm(-x_1_out_00 %*% beta_1_out)
 p_zy1_Sn1_00= pnorm(-x_1_rep_00 %*% beta_1_rep)
 p_zy2_Sn2_00= pnorm(-x_2_rep_00 %*% beta_2_rep)
 p_zy1_Sp1_00= pnorm(-x_1_rep_00 %*% beta_11_rep)
 p_zy2_Sp2_00= pnorm(-x_2_rep_00 %*% beta_22_rep)
 p_ystar_00= p_zy1_Sn1_00 * p_zy2_Sn2_00 * (1-p_zd_00)/ (p_zy1_Sn1_00 * p_zy2_Sn2_00 * (1-p_zd_00) + (1-p_zy1_Sp1_00) * (1-p_zy2_Sp2_00) * p_zd_00);

 # p_ystar_01: probability of ystar is 1 given  y1 is zero and y2 is one (P(Y*=1|Y1=0, Y2=1, X).
 p_zd_01= pnorm(-x_1_out_01 %*% beta_1_out)
 p_zy1_Sn1_01= pnorm(-x_1_rep_01 %*% beta_1_rep)
 p_zy2_Sn2_01= pnorm(-x_2_rep_01 %*% beta_2_rep)
 p_zy1_Sp1_01= pnorm(-x_1_rep_01 %*% beta_11_rep)
 p_zy2_Sp2_01= pnorm(-x_2_rep_01 %*% beta_22_rep)
 p_ystar_01= p_zy1_Sn1_01 * (1-p_zy2_Sn2_01) * (1-p_zd_01)/ (p_zy1_Sn1_01 * (1-p_zy2_Sn2_01) * (1-p_zd_01) + (1-p_zy1_Sp1_01)* p_zy2_Sp2_01 * p_zd_01);

 # p_ystar_10: probability of ystar is 1 given  y1 is one and y2 is zero (P(Y*=1|Y1=1, Y2=0, X).
 p_zd_10= pnorm(-x_1_out_10 %*% beta_1_out)
 p_zy1_Sn1_10= pnorm(-x_1_rep_10 %*% beta_1_rep)
 p_zy2_Sn2_10= pnorm(-x_2_rep_10 %*% beta_2_rep)
 p_zy1_Sp1_10= pnorm(-x_1_rep_10 %*% beta_11_rep)
 p_zy2_Sp2_10= pnorm(-x_2_rep_10 %*% beta_22_rep)
 p_ystar_10= (1-p_zy1_Sn1_10) * p_zy2_Sn2_10 * (1-p_zd_10)/ ((1-p_zy1_Sn1_10) * p_zy2_Sn2_10 * (1-p_zd_10) + p_zy1_Sp1_10 * (1-p_zy2_Sp2_10) * p_zd_10);

 # p_ystar_11: probability of ystar is 1 given  y1 is one and y2 is one (P(Y*=1|Y1=1, Y2=1, X).
 p_zd_11= pnorm(-x_1_out_11 %*% beta_1_out)
 p_zy1_Sn1_11= pnorm(-x_1_rep_11 %*% beta_1_rep)
 p_zy2_Sn2_11= pnorm(-x_2_rep_11 %*% beta_2_rep)
 p_zy1_Sp1_11= pnorm(-x_1_rep_11 %*% beta_11_rep)
 p_zy2_Sp2_11= pnorm(-x_2_rep_11 %*% beta_22_rep)
 p_ystar_11= (1-p_zy1_Sn1_11) * (1-p_zy2_Sn2_11) * (1-p_zd_11)/ ((1-p_zy1_Sn1_11) * (1-p_zy2_Sn2_11) * (1-p_zd_11) + p_zy1_Sp1_11 * p_zy2_Sp2_11 * p_zd_11);
 # Fill in y_star's using the initial guess of the parameters.
 ystar=rep(1, n_total);
 ystar[AHR==0 & EMR==0]= rbinom(n_total_00, size=1, p=p_ystar_00);
 ystar[AHR==0 & EMR==1]= rbinom(n_total_01, size=1, p=p_ystar_01);
 ystar[AHR==1 & EMR==0]= rbinom(n_total_10, size=1, p=p_ystar_10);
 ystar[AHR==1 & EMR==1]= rbinom(n_total_11, size=1, p=p_ystar_11);

 Prev_ystar<- sum(ystar)/N; Prev_ystar

# Subset the covariates using the updated draws, ystar. This step is needed for the DA algorithm
 x_1_out_1x= x_1_out[ystar==1,];
 x_1_out_0x= x_1_out[ystar==0,];
 n_1x= nrow(rbind(x_1_out_1x));
 n_0x= nrow(rbind(x_1_out_0x));

 # Given ystar=1, extract covariates by AHRs
 x_1_rep_1x_1x= x_1_rep[ystar==1 & AHR==1,];
 x_1_rep_1x_0x= x_1_rep[ystar==1 & AHR==0,];
 x_1_rep_1x_obs= rbind(x_1_rep_1x_1x, x_1_rep_1x_0x);
 n_1_rep_1x_1x= nrow(rbind(x_1_rep_1x_1x));
 n_1_rep_1x_0x= nrow(rbind(x_1_rep_1x_0x));
 n_ystar1_y1_obs= nrow(x_1_rep_1x_obs);

 # Given ystar=1, extract covariates by EMRs
 x_2_rep_1x_1x= x_2_rep[ystar==1 & EMR==1,];
 x_2_rep_1x_0x= x_2_rep[ystar==1 & EMR==0,];
 x_2_rep_1x_obs= rbind(x_2_rep_1x_1x, x_2_rep_1x_0x);
  n_2_rep_1x_1x= nrow(rbind(x_2_rep_1x_1x));
 n_2_rep_1x_0x= nrow(rbind(x_2_rep_1x_0x));
 n_ystar1_y2_obs= nrow(x_2_rep_1x_obs);


 ##Set the parameters for the Gibbs chain
 iter_no= 1000; # the number of iterations.
```

```
   #Save the beta parameters draws
  out_para_matrix= matrix(0, nrow= iter_no, ncol= p_cov);  #For the outcome model (beta_O).
  Y1_para_matrix= matrix(0, nrow= iter_no, ncol= q_1_cov); #For the AHR model (beta_Y1).
  Y2_para_matrix= matrix(0, nrow= iter_no, ncol= q_2_cov); #For the EMR model (beta_Y2).

  # Save Y_star for each iteration.
  true_com= matrix(NA, n_total, iter_no)

###########################
##  Begin the DA algorithm  ###
###########################
  cat("begin the cycle", "\n");
  for (iter in 1:iter_no)
  {
    # Step 1: Draw z_O, the truncated normal latent variables for the outcome model.
    z_star_g_1x= rtrun(mu= x_1_out_1x %*% beta_1_out, sigma= rep(1,n_1x), a= rep(0,n_1x),
b= rep(Inf,n_1x));
    z_star_g_0x= rtrun(mu= x_1_out_0x %*% beta_1_out, sigma= rep(1,n_0x), a= rep(-
Inf,n_0x), b= rep(0,n_0x));
     z_star_g_vec= rep(NA, n_total);
    z_star_g_vec[ystar==1]= z_star_g_1x;
    z_star_g_vec[ystar==0]= z_star_g_0x;

    # Step 2: Draw z_Y1 and z_Y2, the truncated normal latent variables for the reporting models.
    z_1_g= c(rtrun(mu= x_1_rep_1x_1x %*% beta_1_rep, sigma= rep(1,n_1_rep_1x_1x), a=
rep(0,n_1_rep_1x_1x), b= rep(Inf,n_1_rep_1x_1x)),
             rtrun(mu= x_1_rep_1x_0x %*% beta_1_rep, sigma= rep(1,n_1_rep_1x_0x), a= rep(-
Inf,n_1_rep_1x_0x), b= rep(0,n_1_rep_1x_0x)));


    z_2_g= c(rtrun(mu= x_2_rep_1x_1x %*% beta_2_rep, sigma= rep(1,n_2_rep_1x_1x), a=
rep(0,n_2_rep_1x_1x), b= rep(Inf,n_2_rep_1x_1x)),
             rtrun(mu= x_2_rep_1x_0x %*% beta_2_rep, sigma= rep(1,n_2_rep_1x_0x), a= rep(-
Inf,n_2_rep_1x_0x), b= rep(0,n_2_rep_1x_0x)));
        # Step 3: Draw beta_O.
    sum_beta_out_g_mean= t(x_1_out) %*% (z_star_g_vec);
    beta_out_g_cov= solve(t(x_1_out) %*% x_1_out);
    beta_out_g_mean= beta_out_g_cov %*% sum_beta_out_g_mean;
    beta_1_out=mvrnorm(n= 1, mu= beta_out_g_mean, Sigma= beta_out_g_cov);

    # Step 4: Draw beta_Y1 and beta_Y2.
    sum_beta_1_rep_mean= t(x_1_rep_1x_obs) %*% (z_1_g);
    beta_1_rep_cov= solve(t(x_1_rep_1x_obs) %*% x_1_rep_1x_obs);
    beta_1_rep_mean= beta_1_rep_cov %*% sum_beta_1_rep_mean;
    beta_1_rep= mvrnorm(n= 1, mu= beta_1_rep_mean, Sigma= beta_1_rep_cov);

    sum_beta_2_rep_mean= t(x_2_rep_1x_obs) %*% (z_2_g );
    beta_2_rep_cov= solve(t(x_2_rep_1x_obs) %*% x_2_rep_1x_obs);
    beta_2_rep_mean= beta_2_rep_cov %*% sum_beta_2_rep_mean;
    beta_2_rep= mvrnorm(n= 1, mu= beta_2_rep_mean, Sigma= beta_2_rep_cov);

    # Save the parameter estimates.
    out_para_matrix[iter,]= beta_1_out;
    Y1_para_matrix[iter,]= beta_1_rep;
    Y2_para_matrix[iter,]= beta_2_rep;

    # Re-calculate the conditional probabilities.
    # p_ystar_00: probability of ystar is 1 given both y1 and y2 are zero P(Y*=1|Y1=0, Y2=0, X).
    p_zd_00= pnorm(-x_1_out_00 %*% beta_1_out)
    p_zy1_Sn1_00= pnorm(-x_1_rep_00 %*% beta_1_rep)
    p_zy2_Sn2_00= pnorm(-x_2_rep_00 %*% beta_2_rep)
    p_zy1_Sp1_00= pnorm(-x_1_rep_00 %*% beta_11_rep)
    p_zy2_Sp2_00= pnorm(-x_2_rep_00 %*% beta_22_rep)
    p_ystar_00= p_zy1_Sn1_00 * p_zy2_Sn2_00 * (1-p_zd_00)/ (p_zy1_Sn1_00 *
  p_zy2_Sn2_00 * (1-p_zd_00) + (1-p_zy1_Sp1_00) * (1-p_zy2_Sp2_00) * p_zd_00);

    # p_ystar_01: probability of ystar is 1 given  y1 is zero and y2 is one P(Y*=1|Y1=0, Y2=1, X).
    p_zd_01= pnorm(-x_1_out_01 %*% beta_1_out)
    p_zy1_Sn1_01= pnorm(-x_1_rep_01 %*% beta_1_rep)
    p_zy2_Sn2_01= pnorm(-x_2_rep_01 %*% beta_2_rep)
    p_zy1_Sp1_01= pnorm(-x_1_rep_01 %*% beta_11_rep)
    p_zy2_Sp2_01= pnorm(-x_2_rep_01 %*% beta_22_rep)
```

```
   p_ystar_01= p_zy1_Sn1_01 * (1-p_zy2_Sn2_01) * (1-p_zd_01)/ (p_zy1_Sn1_01 * (1-
p_zy2_Sn2_01) * (1-p_zd_01) + (1-p_zy1_Sp1_01)* p_zy2_Sp2_01 * p_zd_01);

 # p_ystar_10: probability of ystar is 1 given  y1 is one and y2 is zero P(Y*=1|Y1=1, Y2=0, X).
   p_zd_10= pnorm(-x_1_out_10 %*% beta_1_out)
   p_zy1_Sn1_10= pnorm(-x_1_rep_10 %*% beta_1_rep)
   p_zy2_Sn2_10= pnorm(-x_2_rep_10 %*% beta_2_rep)
   p_zy1_Sp1_10= pnorm(-x_1_rep_10 %*% beta_11_rep)
   p_zy2_Sp2_10= pnorm(-x_2_rep_10 %*% beta_22_rep)
   p_ystar_10= (1-p_zy1_Sn1_10) * p_zy2_Sn2_10 * (1-p_zd_10)/ ((1-p_zy1_Sn1_10) *
p_zy2_Sn2_10 * (1-p_zd_10) + p_zy1_Sp1_10 * (1-p_zy2_Sp2_10) * p_zd_10);

   # p_ystar_11: probability of ystar is 1 given  y1 is one and y2 is one P(Y*=1|Y1=1, Y2=1, X).
   p_zd_11= pnorm(-x_1_out_11 %*% beta_1_out)
   p_zy1_Sn1_11= pnorm(-x_1_rep_11 %*% beta_1_rep)
   p_zy2_Sn2_11= pnorm(-x_2_rep_11 %*% beta_2_rep)
   p_zy1_Sp1_11= pnorm(-x_1_rep_11 %*% beta_11_rep)
   p_zy2_Sp2_11= pnorm(-x_2_rep_11 %*% beta_22_rep)
   p_ystar_11= (1-p_zy1_Sn1_11) * (1-p_zy2_Sn2_11) * (1-p_zd_11)/ ((1-p_zy1_Sn1_11) * (1-
p_zy2_Sn2_11) * (1-p_zd_11) + p_zy1_Sp1_11 * p_zy2_Sp2_11 * p_zd_11);

   # Update the draws of ystar.
   ystar=rep(1, n_total);
   true_m0s0= ystar[AHR==0  & EMR==0 ]= rbinom(n_total_00, size=1, p=p_ystar_00);
   true_m0s1= ystar[AHR==0  & EMR==1 ]= rbinom(n_total_01, size=1, p=p_ystar_01);
   true_m1s0= ystar[AHR==1  & EMR==0 ]= rbinom(n_total_10, size=1, p=p_ystar_10);
   true_m1s1= ystar[AHR==1  & EMR==1 ]= rbinom(n_total_11, size=1, p=p_ystar_11);
  true_com[,iter]= ystar;

   #Subset the covariate matrix using the ystar draws.
   x_1_out_1x= x_1_out[ystar==1,]; n_1x=nrow(rbind(x_1_out_1x));
   x_1_out_0x= x_1_out[ystar==0,]; n_0x=nrow(rbind(x_1_out_0x));

   x_1_rep_1x_1x= x_1_rep[ystar==1 & AHR==1 ,];
n_1_rep_1x_1x=nrow(rbind(x_1_rep_1x_1x));
```

```
    x_1_rep_1x_0x= x_1_rep[ystar==1 & AHR==0 ,];
n_1_rep_1x_0x=nrow(rbind(x_1_rep_1x_0x));
    x_1_rep_1x_obs=rbind(x_1_rep_1x_1x, x_1_rep_1x_0x);
n_ystar1_y1_obs=nrow(x_1_rep_1x_obs);


    x_2_rep_1x_1x= x_2_rep[ystar==1 & EMR==1 ,];
n_2_rep_1x_1x=nrow(rbind(x_2_rep_1x_1x));
    x_2_rep_1x_0x= x_2_rep[ystar==1 & EMR==0 ,];
n_2_rep_1x_0x=nrow(rbind(x_2_rep_1x_0x));
    x_2_rep_1x_obs=rbind(x_2_rep_1x_1x, x_2_rep_1x_0x);
n_ystar1_y2_obs=nrow(x_2_rep_1x_obs);
  }
  ############ End of the Gibbs chain ############


#################
### Final PSSA ###
 ################
 P_PSSA= sum(ystar)/N;
Prevalence = cbind(P_True, P_AHR, P_EMR, T_correlation, P_correlation, P_OR, P_AND,
P_RSSA, P_PSSA)


  return(Prevalence)
}


Performance= function(N, condition=c(), P_True){
  # N= sample size


  ### Calculate sample mean, variance and standard error of the prevalence estimates ###
  condition= condition
  Mean_AHR= mean(condition[,2,]); Mean_EMR= mean(condition[,3,]);
  Mean_T_correlation= mean(condition[,4,]); Mean_P_correlation= mean(condition[,5,]);
  Mean_OR= mean(condition[,6,]);  Mean_AND= mean(condition[,7,]);
  Mean_RSSA= mean(condition[,8,]); Mean_PSSA= mean(condition[,9,])
  Variance_OR= var(condition[,6,]);   Variance_AND= var(condition[,7,]);
  Variance_RSSA= var(condition[,8,]); Variance_PSSA= var(condition[,9,])
  SE_AHR= sqrt(var(condition[,2,])/ N); SE_EMR= sqrt(var(condition[,3,])/ N);  SE_OR=
sqrt(var(condition[,6,])/ N);
  SE_AND= sqrt(var(condition[,7,])/ N); SE_RSSA= sqrt(var(condition[,8,])/ N); SE_PSSA=
sqrt(var(condition[,9,])/ N)

  #### Performance Measures ###
  P_True= P_True
  Bias_OR= abs(P_True - Mean_OR); Bias_AND= abs(P_True - Mean_AND); Bias_RSSA=
abs(P_True - Mean_RSSA); Bias_PSSA= abs(P_True - Mean_PSSA)

  Rel.Bias_OR= abs(P_True - Mean_OR)/ P_True; Rel.Bias_AND= abs(P_True - Mean_AND)/
P_True; Rel.Bias_RSSA= abs(P_True - Mean_RSSA)/ P_True
  Rel.Bias_PSSA= abs(P_True - Mean_PSSA)/ P_True

  MSE_OR= Variance_OR + Bias_OR^2; MSE_AND= Variance_AND + Bias_AND^2;
MSE_RSSA= Variance_RSSA + Bias_RSSA^2; MSE_PSSA= Variance_PSSA + Bias_PSSA^2

  Measures = cbind(Mean_AHR, Mean_EMR, Mean_T_correlation, Mean_P_correlation,
                   Mean_OR, Mean_AND, Mean_RSSA, Mean_PSSA,
                   SE_AHR, SE_EMR, SE_OR, SE_AND, SE_RSSA, SE_PSSA,
                   Bias_OR, Bias_AND, Bias_RSSA, Bias_PSSA,
                   Rel.Bias_OR, Rel.Bias_AND, Rel.Bias_RSSA, Rel.Bias_PSSA,
                   MSE_OR, MSE_AND, MSE_RSSA, MSE_PSSA)
   return(Measures)

}

##############################
### Simulation parameters ######
##############################
sim_no= 500
N= 10000
Sn1= c(0.70, 0.70, 0.72, 0.80, 0.65, 0.82, 0.68, 0.71);
Sp1= c(0.95, 0.99, 0.96, 0.98, 0.94, 0.99, 0.97, 0.985)
Sn2= c(0.75, 0.65, 0.55, 0.65, 0.50, 0.50, 0.55, 0.71);
Sp2= c(0.99, 0.99, 0.95, 0.98, 0.99, 0.99, 0.95, 0.985)
```

```
xi_dist= c("ex","un")
xi_corr_ex= c(0.00, 0.20, 0.50)
n_xi= c(8,16)
xi_mean= c(0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.50, 0.50, 0.50, 0.50, 0.30, 0.30, 0.70,
0.70)
#xi_mean= c(0.05, 0.10, 0.10, 0.05, 0.15, 0.10, 0.10, 0.15, 0.05, 0.15, 0.15, 0.20, 0.15, 0.20,
0.10, 0.20)
p10_xi_betas8=  c(-0.35, 1.10, 0.20, 0.50, 0.15, 0.05, -0.50, -1.30, -1.80, 0,0,0,0,0,0,0,0)
p20_xi_betas8=  c(-0.35, 1.20, 0.90, 0.50, 0.15, 0.05, -0.20, -0.80, -1.30, 0,0,0,0,0,0,0,0)
p10_xi_betas16= c(-0.35, 1.10, 0.20, 0.40, 0.15, 0.02, -0.60, -0.50, -1.80, 0.02, 0.02, 0.02, 0.02,
0.50, 0.50, -0.50, -0.50)
p20_xi_betas16= c(-0.35, 1.10, 0.80, 0.40, 0.15, 0.04, -0.20, -0.35, -1.45, 0.04, 0.04, 0.04, 0.04,
0.40, 0.40, -0.35, -0.35)
colnames<- list(c("P_True", "P_AHR", "P_EMR", "T_Corr", "P_Corr", "P_OR", "P_AND",
"P_RSSA", "P_PSSA"))

## Generate 100 seeds via a random process and save them for reproducibility
ranseed <- round(runif(100)*1000000)

# To test any of the simulation conditions replace the characters in the brackets for R function
“thesis” with the appropriate simulation parameter or vector of parameters. For example:
Condition 1
set.seed(ranseed[1])
condition1= replicate(sim_no, thesis(N, Sn1[3], Sp1[3], Sn2[3], Sp2[3], n_xi[2], xi_corr_ex[1],
xi_dist[1], xi_mean, p20_xi_betas16))
P_True1= 0.20   # Fix the true prevalence
Measure1= Performance(N, condition1, P_True1)

# Final output 1
Final_output1= data.frame("seed no"= ranseed[1],"N"= N, Sn1[3], Sp1[3], Sn2[3], Sp2[3],
n_xi[2], xi_corr_ex[1], xi_dist[1], P_True1, Measure1)
```

APPENDIX C: Descriptive Statistics for Case Ascertainment Markers for the PSSA Method

Table C.1: Frequency and percentage of case ascertainment markers from administrative health records

	Overall		**Males**		**Females**		**18-44 years**		**45-64 years**		**65+ years**	
Condition	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
CD	916	1.3	376	1.3	540	1.4	72	0.2	366	1.4	478	5.0
CHD	2623	3.8	1553	5.2	1070	2.7	129	0.4	1109	4.2	1385	14.4
CHF	558	0.8	248	0.8	310	0.8	29	0.1	177	0.7	352	3.7
COPD	1287	1.9	604	2.0	683	1.7	137	0.4	557	2.1	593	6.2
Dementia	625	0.9	259	0.9	366	0.9	206	0.6	196	0.7	223	2.3
Depression	7098	10.3	2013	6.8	5085	13.0	3558	10.8	2827	10.8	713	7.4
Diabetes	4176	6.1	2003	6.7	2173	5.6	701	2.1	2134	8.1	1341	13.9
Obesity	1623	2.4	536	1.8	1087	2.8	772	2.3	710	2.7	141	1.5
RD	916	1.3	471	1.6	445	1.1	259	0.8	428	1.6	229	2.4
SA	1387	2.0	716	2.4	671	1.7	811	2.5	501	1.9	75	0.8

Note: COPD= Chronic obstructive pulmonary disease; CD= Cerebrovascular disease; CHF= Congestive heart failure; CHD= Coronary heart disease; RD= Renal disease; SA= Substance abuse

Table C.2: Frequency and percentage of case ascertainment markers identified electronic medical records

	Overall		**Males**		**Females**		**18-44 years**		**45-64 years**		**65+ years**	
Condition	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
COPD	181	0.3	97	0.3	84	0.2	43	0.1	75	0.3	63	0.7
Dementia	1130	1.6	408	1.4	722	1.8	225	0.7	282	1.1	623	6.5
Depression	11005	16.0	3428	11.5	7577	19.4	5544	16.8	4168	15.9	1293	13.4
Diabetes	6435	9.3	3194	10.7	3241	8.3	1491	4.5	3288	12.5	1656	17.2
Obesity	15191	22.1	6889	23.1	8302	21.2	6171	18.7	6951	26.5	2069	21.5

Note: COPD= Chronic obstructive pulmonary disease

Table C.3: Polychoric and tetrachoric correlations for case ascertainment markers

	Region	**Sex**	**Age group**	**CCS**	**Q**	**DB$_{A}$**	**CHD$_{A}$**	**OB$_{A}$**	**COPD$_{A}$**	**CD$_{A}$**	**CHF$_{A}$**	**DP$_{A}$**	**DM$_{A}$**	**RD$_{A}$**	**SA$_{A}$**	**DB$_{E}$**	**COPD$_{E}$**	**DP$_{E}$**	**DM$_{E}$**	**OB$_{E}$**
Region	1.00	-0.02	0.13	0.16	0.02	0.12	0.12	-0.02	0.11	0.16	0.10	0.10	0.11	0.04	0.10	-0.05	-0.58	-0.13	0.09	0.31
Sex		1.00	0.00	0.03	-0.07	-0.06	-0.18	0.11	-0.04	0.02	-0.01	0.22	0.02	-0.08	-0.09	-0.09	-0.08	0.21	0.07	-0.04
Age group			1.00	0.34	-0.07	0.38	0.56	-0.03	0.42	0.45	0.50	-0.06	0.20	0.18	-0.13	0.31	0.22	-0.05	0.40	-0.02
CCS				1.00	-0.10	0.78	0.39	0.12	0.54	0.58	0.60	0.13	0.42	0.37	0.15	0.46	0.14	0.05	0.25	0.14
Q					1.00	-0.12	-0.08	-0.01	-0.12	-0.08	-0.13	-0.07	-0.17	-0.02	-0.11	-0.07	-0.16	-0.05	-0.13	-0.04
DB$_{A}$						1.00	0.33	0.16	0.19	0.24	0.35	0.07	0.08	0.28	0.02	0.80	0.07	0.03	0.14	0.17
CHD$_{A}$							1.00	0.08	0.30	0.41	0.61	0.02	0.20	0.22	0.00	0.27	0.02	-0.02	0.26	0.05
OB$_{A}$								1.00	0.02	0.06	0.18	0.19	-0.01	0.04	0.05	0.22	0.02	0.16	-0.04	0.17
COPD$_{A}$									1.00	0.18	0.39	0.09	0.17	0.14	0.21	0.16	0.50	0.05	0.17	0.07
CD$_{A}$										1.00	0.36	0.10	0.36	0.20	0.13	0.16	-1.00	0.00	0.30	0.07
CHF$_{A}$											1.00	0.06	0.29	0.33	0.08	0.25	0.19	-0.01	0.28	0.13
DP$_{A}$												1.00	0.40	0.09	0.32	0.03	0.08	0.57	0.27	0.10
DM$_{A}$													1.00	0.11	0.60	0.02	0.12	0.25	0.68	0.14
RD$_{A}$														1.00	0.06	0.18	-0.02	0.06	0.02	0.05
SA$_{A}$															1.00	-0.05	0.04	0.11	0.04	0.12
DB$_{E}$																1.00	0.18	0.14	0.14	-0.05
COPD$_{E}$																	1.00	0.26	0.18	-0.20
DP$_{E}$																		1.00	0.38	-0.28
DM$_{E}$																			1.00	-0.03
OB$_{E}$																				1.00

Note: CCS= Charlson Comorbidity Score; Q= Income quintile; DB= Diabetes; CHD= Coronary heart disease; OB= Obesity; COPD= Chronic obstructive pulmonary disease; CD= Cerebrovascular disease; CHF= Congestive heart failure; DP= Depression; DM= Dementia; RD= Renal disease; SA= Substance abuse; Subscripts A and E denote whether the marker was identified from AHRs or EMRs, respectively

APPENDIX D: Diagnostics plots of the Posterior Distribution of the Estimated Disease Prevalence for the PSSA Method

Figure D.1: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, overall

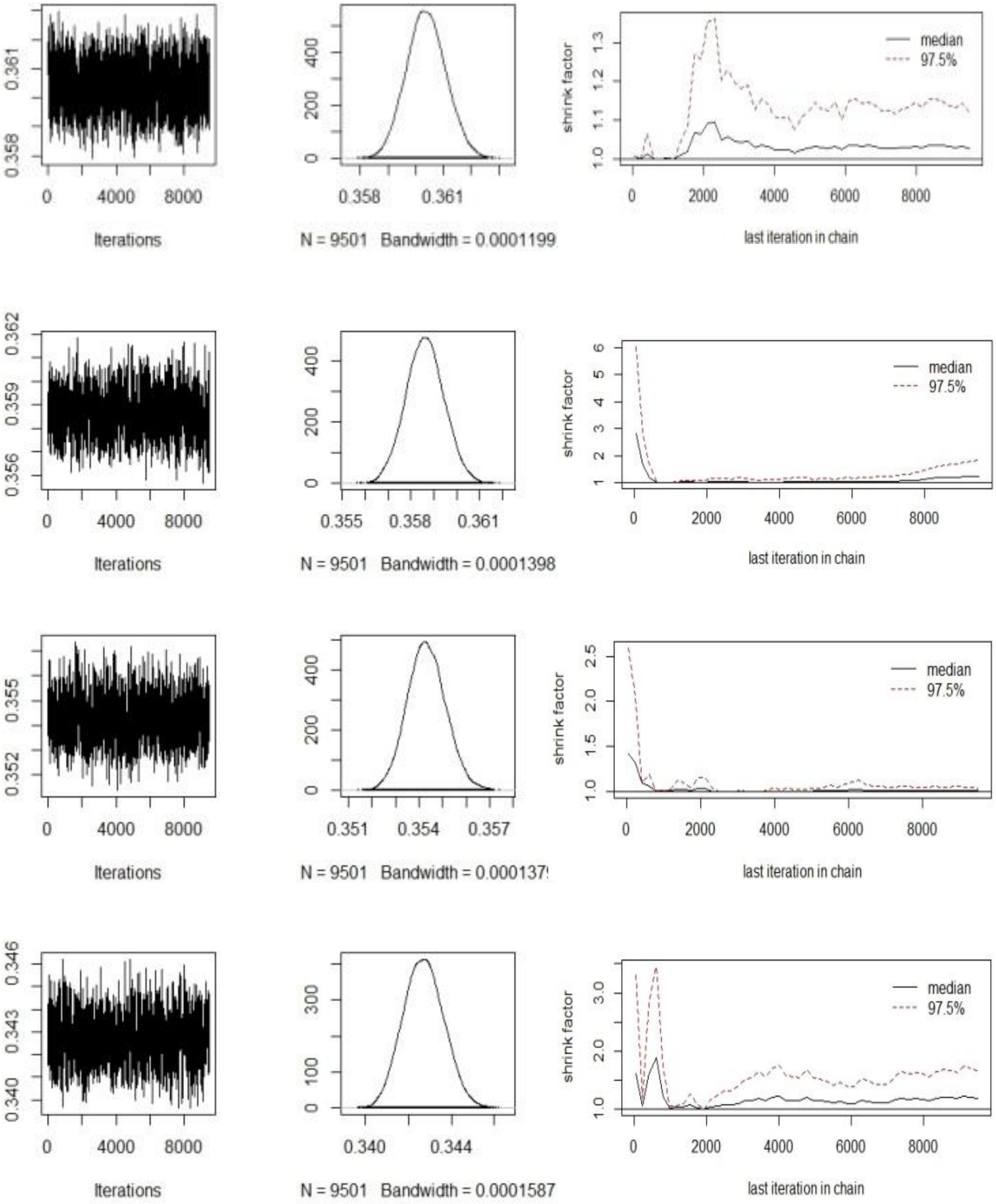


Figure D.2: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, males

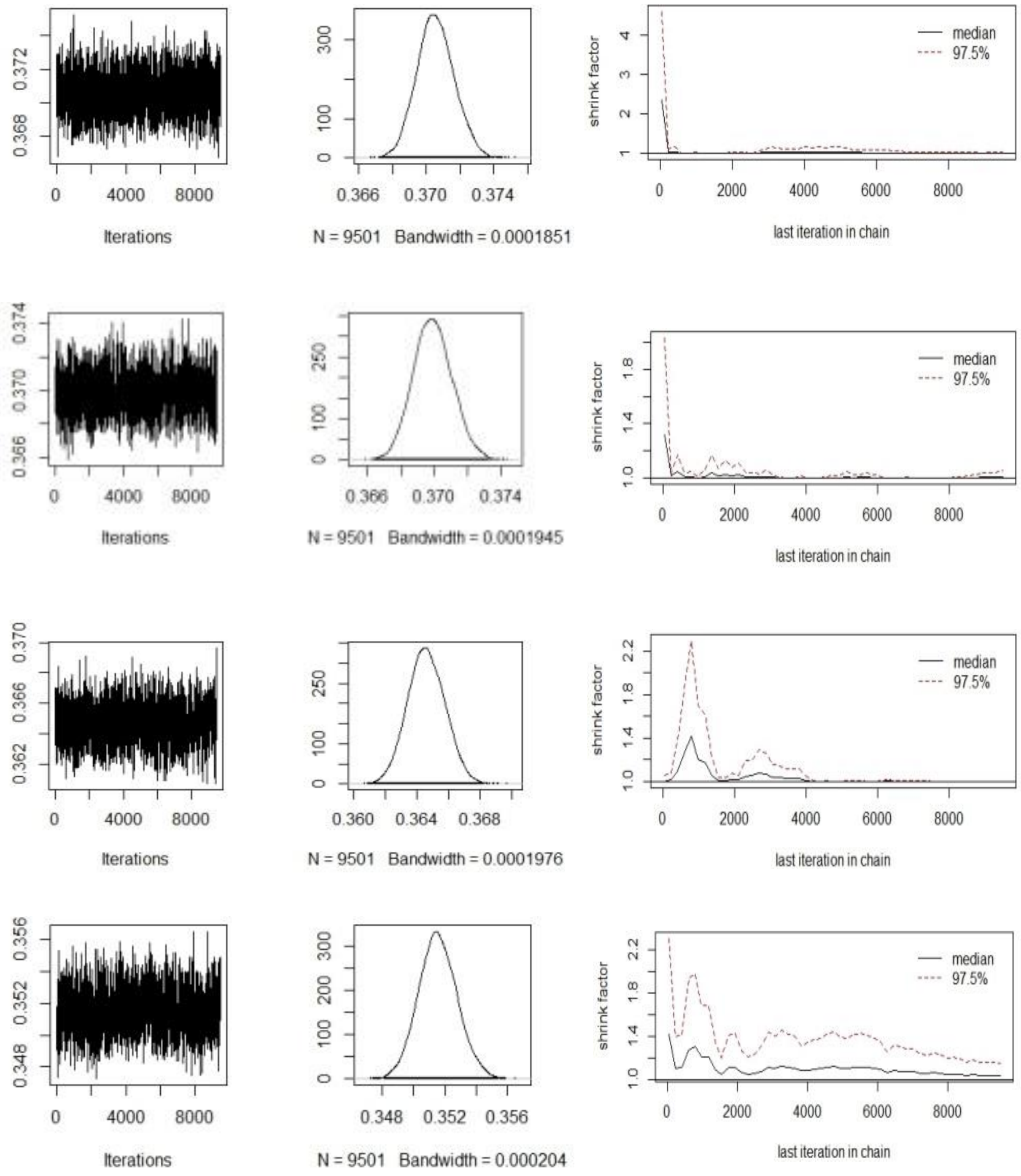


Figure D.3: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, females

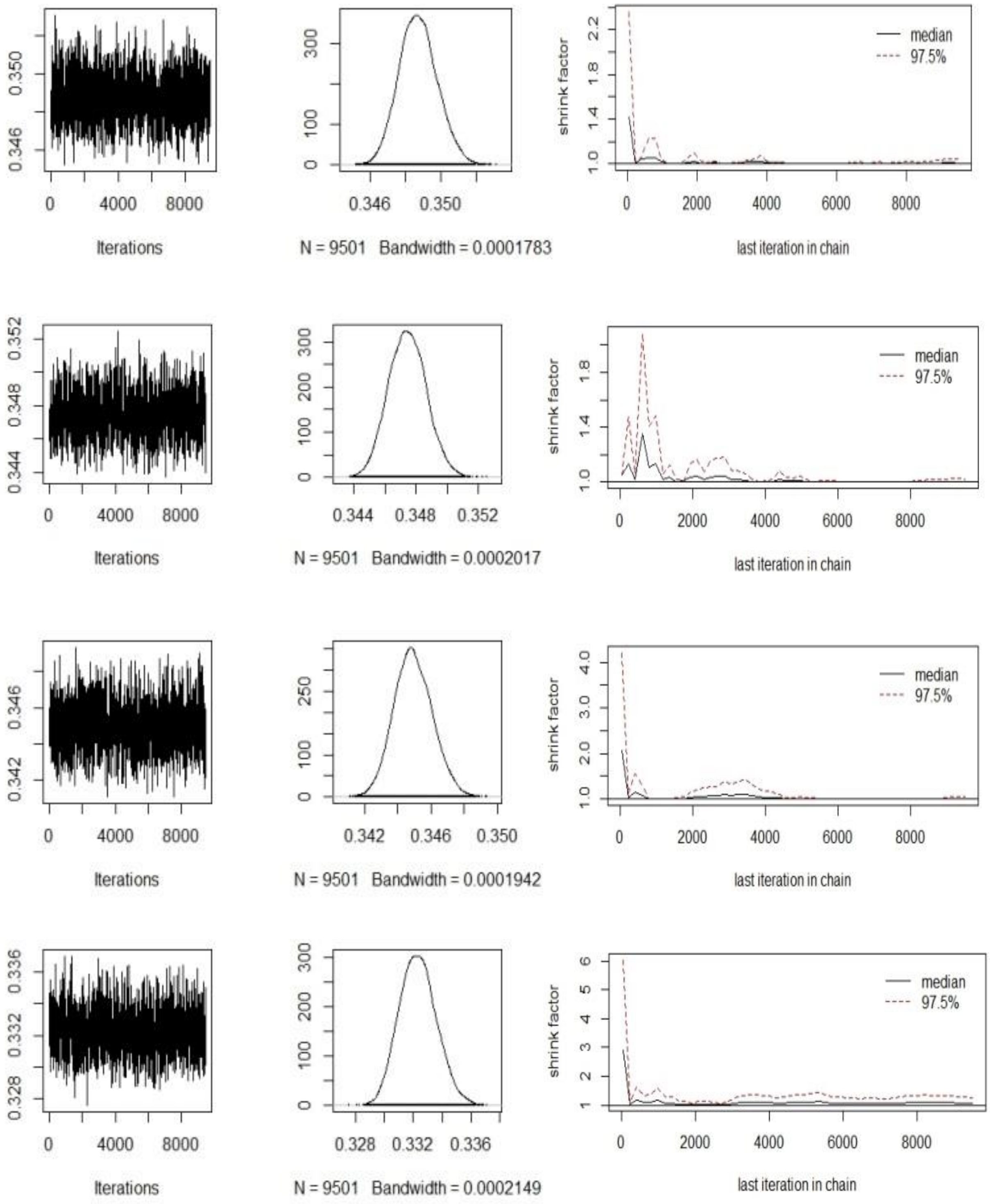


Figure D.4: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, 18-44 age group

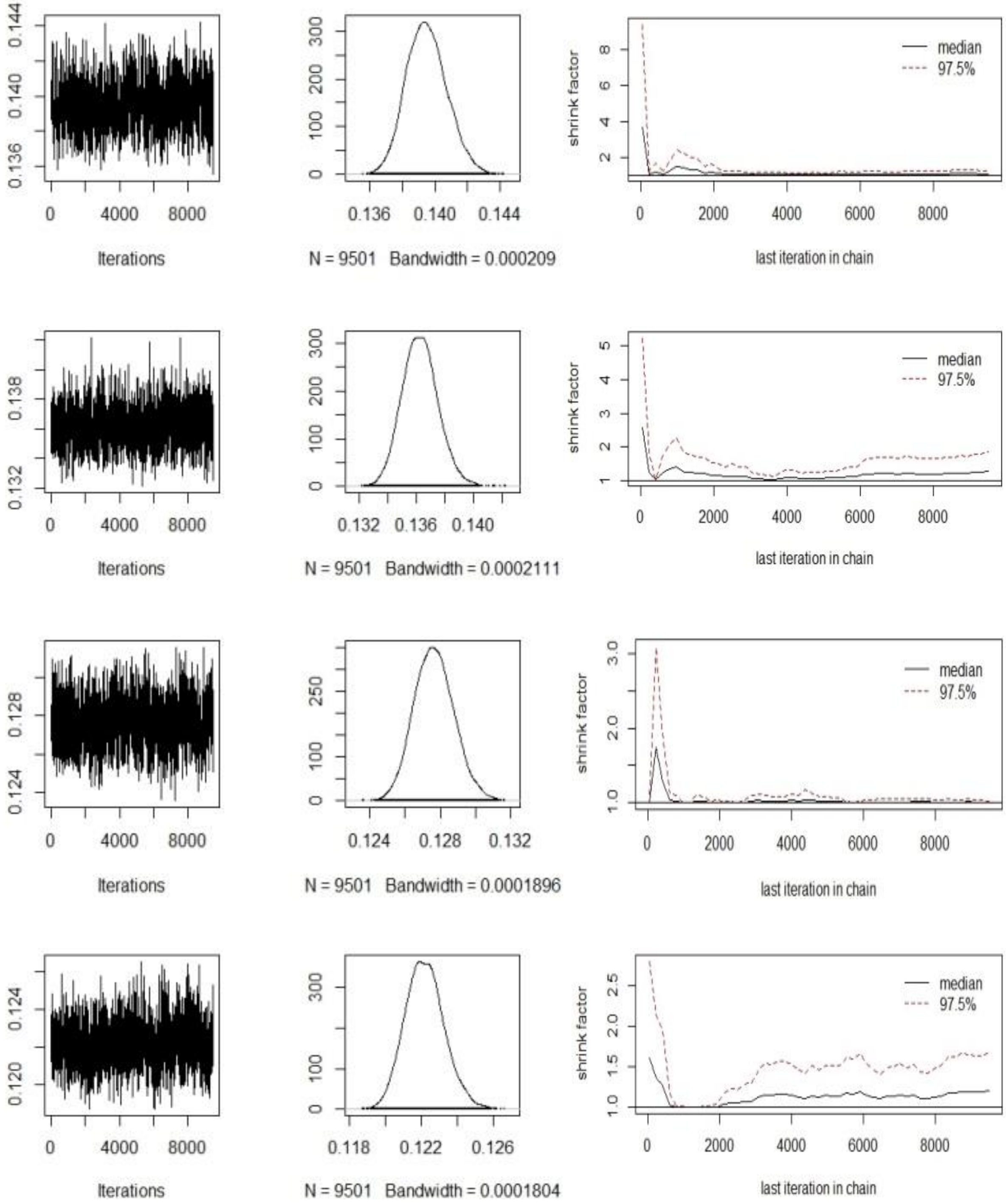


Figure D.5: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, 45-64 age group

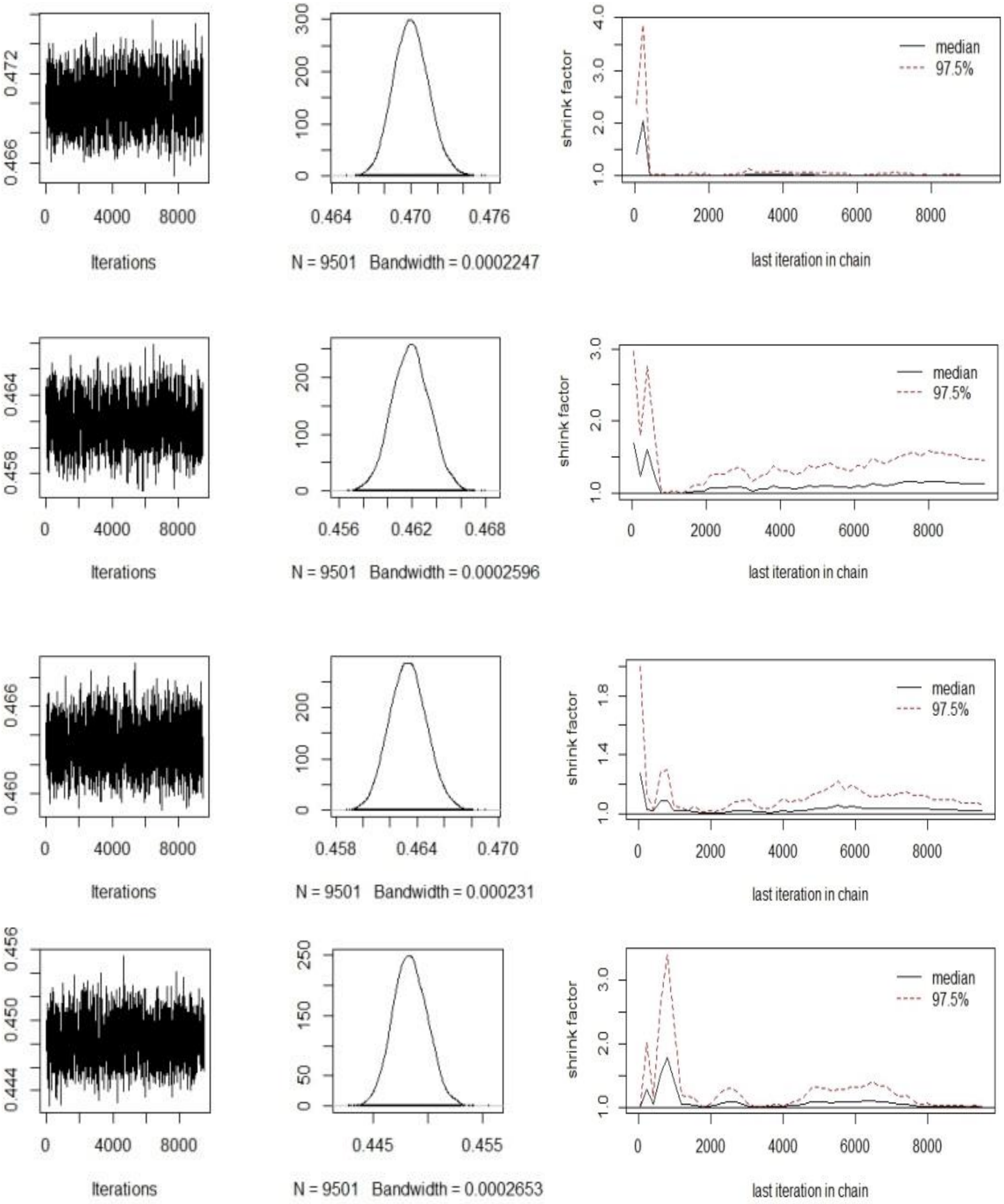
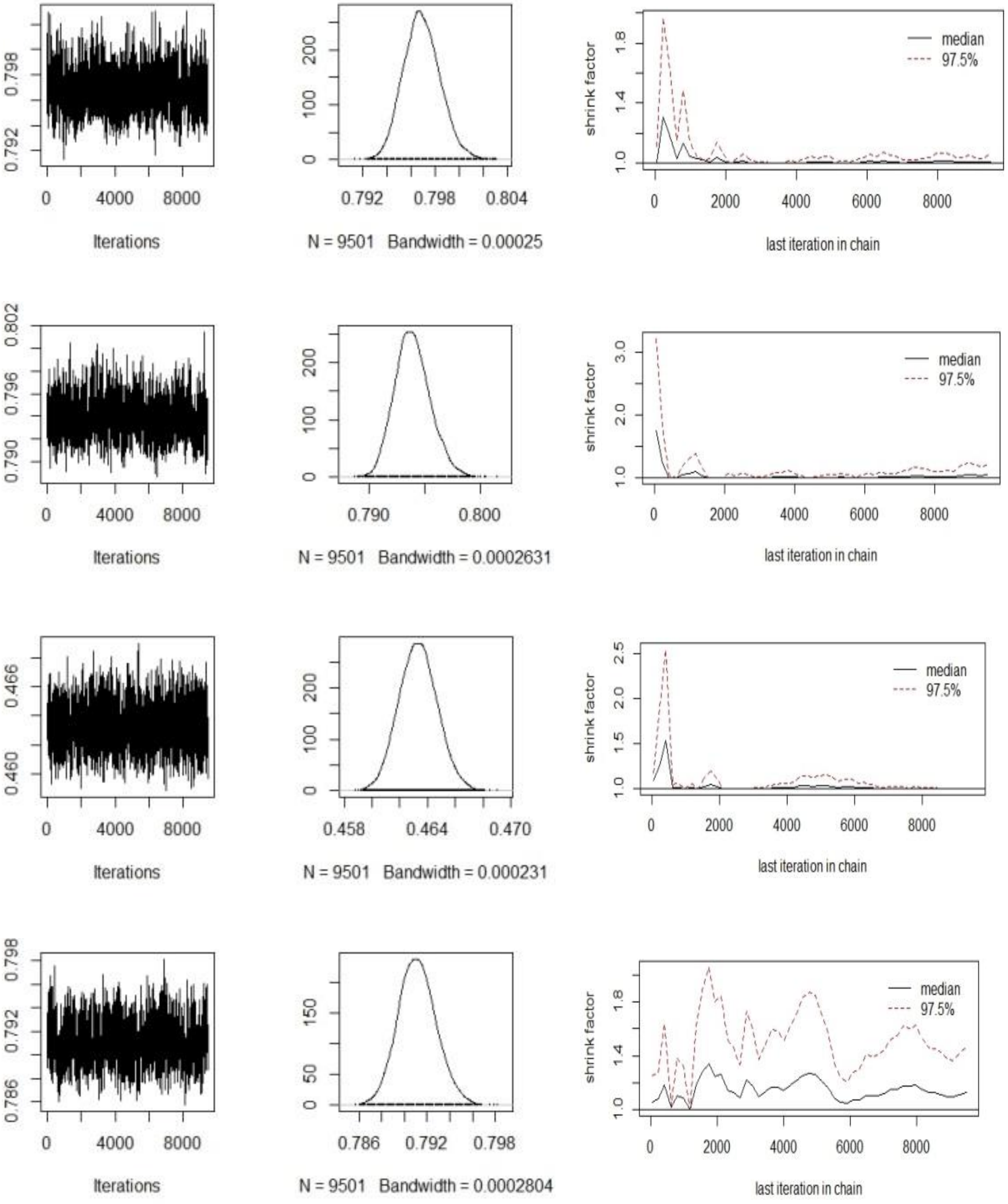


Figure D.6: Trace plots, density plots and convergence plots of the posterior distribution of the estimated disease prevalence for the PSSA method, 65+ age group


Iterations
N = 9501 Bandwidth = 0.00025
shrink factor
median
last iteration in chain
Iterations
N = 9501 Bandwidth = 0.0002631
shrink factor
median
last iteration in chain
Iterations
N = 9501 Bandwidth = 0.000231
shrink factor
median
last iteration in chain
Iterations
N = 9501 Bandwidth = 0.0002804
shrink factor
median
last iteration in chain