Modeling Diagnostic Validity Estimates from Administrative Health Data: Application to Rheumatoid Arthritis

by

Kristine Kroeker

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 data are known to contain measurement error in their diagnostic information, particularly for chronic diseases. Diagnostic validation studies are used to assess the accuracy of administrative health data and produce estimates such as sensitivity and specificity, which are typically used to recommend an optimal case definition(s) for cohort selection. Currently, many researchers test multiple case definitions in a single study and rely on descriptive analyses to select one or more case definitions for use in practice.

Purpose & Objectives: The research purpose was to develop and assess model-based methods to select a case definition for identifying individuals with a chronic disease in administrative health data. The objectives were to: (1) compare the performance of three regression models to test for differences in the accuracy of case definitions, and (2) demonstrate how to apply and use these models in a real numeric example.

Methods: A simulation study was used to compare the performance of the following regression models: (a) univariate models independently applied to sensitivity and specificity estimates, (b) bivariate model applied to estimates of sensitivity and specificity, and (c) univariate model applied to a summary measure of sensitivity and specificity. Model comparisons were done using the number of simulations that converge, bias, mean squared error (MSE), and 95% confidence interval (CI) coverage. The simulation study investigated five parameters. The models were demonstrated using secondary analysis of 148 case definitions from a rheumatoid arthritis (RA) diagnostic validation study.

Results: The univariate models had low bias, low MSE, reasonable 95% CI coverage, and converged in all scenarios. The bivariate model had low bias, low MSE, poor 95% CI coverage, and converged in 81% of the scenarios. The RA case definition characteristics that showed to be

positively associated with sensitivity or specificity were requiring at least two physician diagnoses, having an unlimited number of years to ascertain cases, requiring at least one specialist diagnosis, and requiring at least one prescription for individuals 65 years of age and older.

Conclusion: The results suggest that these models provide researchers with an inferential method for identifying the case definition characteristics associated with the validity measures.

ACKNOWLEDGEMENTS

I would like to start by thanking my supervisor, Dr. Lisa Lix, for all of her guidance and support through the Master's program. She has provided me with invaluable insight into my thesis project and helped me improve my skills as a researcher. Thank you to my committee members, Dr. Saman Muthukumarana and Dr. Depeng Jiang, who have provided constructive feedback on my thesis. Thank you to Dr. Jessica Widdifield who provided me with another perspective on my discussion. My thesis would not have been as complete without all of the feedback I received.

I appreciate the work space the George and Fay Yee Centre for Healthcare Innovation had provided me. I am grateful for the financial support I received from the Canadian Institutes of Health Research (CIHR) and the Canadian Network for Advanced Interdisciplinary Methods for comparative effectiveness research (CAN-AIM) to complete my thesis in a timely manner.

Thank you to my family for providing me with support to achieve my academic goals. I am grateful to my husband, Dean Kroeker, who provided me with endless support and encouragement. Thank you for putting up with my busy schedule and listening to me talk about my thesis even when you didn't understand.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Abbreviation	Definition
AIC	Akaike information criterion
ANOVA	Analysis of variance
BIC	Bayesian information criterion
CDF	Cumulative distribution function
CI	Confidence interval
DMARD	Disease-modifying antirheumatic drugs
DOR	Diagnostic odds ratio
EMERALD	Electronic Medical Records Administrative Linked Database
ER	Emergency room
FN	False negative
FP	False positive
ICD-9	International Classification of Diseases, 9^{th} revision
ICD-10	International Classification of Diseases, 10^{th} revision
IML	Interactive matrix language
JRA	Juvenile rheumatoid arthritis
LRT	Likelihood ratio test
MSE	Mean square error
NPV	Negative predictive value
PPV	Positive predictive value
QUADAS	Quality Assessment of Diagnostic Accuracy Studies
R^2	Percent of explained variance
RA	Rheumatoid arthritis
REML	Restricted maximum likelihood
RX	Prescription medication
SD	Standard deviation
SE	Standard error
STARD	Standards for Reporting of Diagnostic Accuracy
TAP	Test to Actual Positives
TN	True negative
TP	True positive

CHAPTER 1 – INTRODUCTION

1.1 Background

In Canada, as in other countries, integrating and linking information from different sources is helping to create rich data repositories for population health research. Electronic administrative health databases, which capture information about patient contacts with the healthcare system, including hospitalizations, physician visits, and prescription drug use, are a central component of these data repositories. These data were originally intended for managing the healthcare system and paying providers. However, administrative health databases are increasingly used for secondary purposes, such as chronic disease research and surveillance (Katz et al., 1997; Leslie, Lix, & Yogendran, 2011; Lix et al., 2006; Quan et al., 2009), post-marketing studies of adverse drug effects (Suissa et al., 2012), and comparative effectiveness studies about healthcare treatments and providers (Fung, Brand, Newhouse, & Hsu, 2011). With respect to the first of these purposes, studies about such diseases as arthritis, diabetes, and hypertension using administrative health databases have greatly improved our understanding of the impact of these conditions on the Canadian population (Bernatsky et al., 2014; Dart et al., 2011; Quan et al., 2009).

Administrative health databases are advantageous for research and surveillance studies because they are relatively inexpensive to access, cover entire populations, and can be linked to create longitudinal patient-specific records of healthcare use. However, one limitation is their quality for research and surveillance, particularly the potential for low sensitivity of diagnoses to identify patients with chronic diseases (Hux, Ivis, Flintoft, & Bica, 2002; Lix et al., 2008; Quan et al., 2009; Tu, Campbell, Chen, Cauch-Dudek, & McAlister, 2007). Diagnostic validation studies are an essential tool for assessing data quality to ensure unbiased research results. These

studies compare diagnosed cases in administrative health data, as defined by a case definition, with clinically-confirmed cases. These studies produce diagnostic validity measures (Lix et al., 2014, 2008) for one or more case definitions, the rules applied to administrative health data to identify individuals in the population with the disease of interest.

Guidelines to evaluate the reporting quality of validation studies (Benchimol et al., 2011) and to evaluate the quality of diagnostic accuracy studies (Whiting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003) have been developed. However, these guidelines have not addressed the issues that arise when multiple case definitions are tested in a single validation study. Multiple definitions are often compared in order to identify an optimal definition(s), usually one that simultaneously maximizes sensitivity (i.e., the proportion of correctly identified positive cases) and specificity (i.e., the proportion of correctly identified negative cases). Selecting the optimal definition may not be a straightforward process, particularly if sensitivity and specificity vary with data characteristics, such as the choice of diagnosis codes, the number of years of data used to ascertain cases, patient characteristics such as age and sex, or over time. Existing guidelines recommend selecting a case definition by prioritizing a validity measure to fit the study purpose to limit the misclassification bias (Chubak, Pocobelli, & Weiss, 2012). Many researchers rely on descriptive analyses to select a case definition; these analyses do not account for sampling error in the estimates. Developing inferential methods to select and recommend the optimal case definition(s) could assist researchers to more accurately identify cases of chronic diseases in administrative health data.

Regression methods have been used in meta-analyses to combine estimates of sensitivity and specificity for one or more case definitions from different diagnostic validation studies. Univariate, bivariate, and even trivariate random-effects models (Chu, Guo, & Zhou, 2010; Diaz,

2015; Hoyer & Kuss, 2015; Kuss, Hoyer, & Solms, 2014; Reitsma et al., 2005) have been used to model estimates of sensitivity and specificity from different studies; the random effects account for between-study variation arising because of dependence in the diagnostic validity estimates from case definitions. The use of meta-analytic regression models parallels our interest in using regression models to test the association of diagnostic validation study characteristics with estimates of the accuracy of different case definitions within a single study.

While diagnostic validation studies have been undertaken for many chronic diseases, the focus of our research was on validation studies for rheumatoid arthritis (RA), an inflammatory condition that results in joint inflammation and pain. Accurate information about the prevalence and incidence of RA is critical for understanding the burden of disease in the population, and for planning disease treatment and management programs. Several studies have been conducted about case definition(s) for administrative health data to ascertain RA in adults (Katz et al., 1997; Kim et al., 2011; Ng, Aslam, Petersen, Yu, & Suarez-Almazor, 2012; Singh, Holmgren, & Noorbaloochi, 2004; Tennis, Bombardier, Malcolm, & Downey, 1993; Widdifield, Bernatsky, et al., 2013; Widdifield et al., 2014). There have also been validation studies about case definitions for children and youth with juvenile RA (JRA) (Harrold et al., 2013; Stringer & Bernatsky, 2015).

1.2 Purpose and Objectives

The research purpose was to develop and assess methods to select a case definition for identifying individuals with chronic disease in administrative health data. The objectives were:

1. To compare three regression models to test for differences in the accuracy of case definitions from diagnostic validation studies: (a) univariate fixed-effects models independently applied to sensitivity and specificity estimates, (b) bivariate mixed-effects

model applied to estimates of sensitivity and specificity, and (c) univariate fixed-effects model applied to a summary measure of sensitivity and specificity.

2. To demonstrate how to apply and use these models in diagnostic validation studies, with a specific application involving RA.

1.3 Thesis Organization

Chapter 2 includes a literature review on four topics: a) the quality of diagnostic validation studies in administrative health data, b) characteristics of diagnostic validity measures, c) regression models for meta-analysis of diagnostic validation studies, and d) rheumatoid arthritis diagnostic validation studies. Chapter 3 contains the methods for the simulation study and the numeric example. Chapter 4 presents the results of the simulation study and the numeric example. The last chapter includes the study discussion, conclusions, and opportunities for future research.

CHAPTER 2 – LITERATURE REVIEW

This literature review encompasses the following topics: quality of diagnostic validation studies in administrative health data, characteristics of diagnostic validity measures, regression models for meta-analysis of diagnostic validation studies, and rheumatoid arthritis diagnostic validation studies.

2.1 Quality of Diagnostic Validation Studies in Administrative Health Data

Criteria that have been developed to evaluate diagnostic validation studies include: (a) the Standards for Reporting of Diagnostic Accuracy (STARD) criteria for evaluating the reporting (Bossuyt et al., 2003) and (b) the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) criteria for evaluating the quality (Whiting et al., 2003). These guidelines have been recommended for use when reporting and evaluating validation studies (Benchimol et al., 2011; Bernatsky, Lix, O'Donnell, Lacaille, & CANRAD Network, 2013; Jolley et al., 2015; Widdifield, Labrecque, et al., 2013).

Benchimol et al.'s (2011) checklist, which is based on the STARD criteria, is used to evaluate the reporting quality of the methods and results of validation studies. The study recommends many reporting guidelines but focuses on the following as the most important for all future validation studies to follow and consider: describing the validation cohort, providing at least four diagnostic validity estimates with confidence intervals (CIs) for each case definition (e,g., sensitivity, specificity, positive predictive value [PPV], and negative predictive value [NPV]), ensuring the prevalence values in the reference standard and administrative health data are equal when estimating predictive values, and validating a case definition in other populations to confirm its accuracy (Benchimol et al., 2011).

2.2 Characteristics of Diagnostic Validity Measures

In a diagnostic validation study, an evaluation of a diagnostic case definition is conducted in the following sequence of steps. First, an individual is identified as having positive or negative disease status in administrative health data using diagnostic information recorded in the data. Second, this status is compared with an individual's true disease status in a reference standard data source. Third, the individual is classified as a true positive (TP), true negative (TN), false positive (FP), or false negative (FN). Lastly, the classifications are used to calculate estimates of diagnostic validity measures, including sensitivity, specificity, PPV, and NPV. Sensitivity and specificity are the most commonly reported measures. Sensitivity is defined as the proportion of individuals identified with the disease given that the individual has the disease (Giavarina, 2012). Specificity is defined as the proportion of individuals identified without the disease given that the individual does not have the disease (Giavarina, 2012). The two measures are negatively correlated (Reitsma et al., 2005). It has been recommended that sensitivity and specificity always be reported together to completely assess the accuracy of a diagnostic test (Jones, Ashrafian, Darzi, & Athanasiou, 2010).

Vecchio (1966) developed PPV and NPV to predict the likelihood of a positive or negative test result given that disease status was positive or negative, respectively. Previous research has shown that when sensitivity and specificity are held constant, PPV will increase as disease prevalence increases (Vecchio, 1966). Additionally, when disease prevalence is held constant, PPV will increase as specificity increases (Vecchio, 1966). Hence, Vecchio (1966) suggested that PPV and NPV be reported in validation studies when sensitivity, specificity, and disease prevalence are known. However, Giavarina et al. (2012) recommended that PPV and

NPV not be reported in diagnostic validation studies because they are dependent on disease prevalence, which is often unknown.

These measures of sensitivity, specificity, PPV, and NPV do not contain sufficient information on their own to describe a case definition's discriminatory power (Glas, Lijmer, Prins, Bonsel, & Bossuyt, 2003). Therefore, the measures have also been combined to create summary measures of test performance including: (a) the diagnostic odds ratio (DOR) (Glas et al., 2003), (b) Youden's (1950) index, and (c) the Test to Actual Positives (TAP) (Campbell, Biloglav, & Rudan, 2008). The DOR is the odds of an individual being correctly identified as disease positive compared to the odds of an individual being incorrectly identified as disease positive:

$$DOR = \frac{Y_{sens} * Y_{spec}}{(1 - Y_{sens}) * \left(1 - Y_{spec}\right)}, \tag{2.1}$$

where Y_{sens} is the value of sensitivity and Y_{spec} is the value of specificity. The DOR ranges from zero to infinity with higher values indicating better accuracy and a value of one indicating the test identifies the same number of individuals as disease positive and negative (Glas et al., 2003). Youden's (1950) index is defined as:

$$W = Y_{sens} + Y_{spec} - 1, \tag{2.2}$$

and theoretically ranges from negative one to positive one with higher values indicating better discriminatory power and values of zero or below indicating the test is meaningless (Chen, Xue, Tan, & Chen, 2015). In practice, Youden's (1950) index is not commonly reported when it is below zero. The DOR and Youden's (1950) index are independent of disease prevalence and therefore do not change if the population's disease prevalence changes (Glas et al., 2003; Youden, 1950).

A disadvantage of using the DOR and Youden's (1950) index is that they can produce the same estimate despite the fact that sensitivity and specificity may have different values. However, this issue could be addressed by adopting a weighted index of sensitivity and specificity, with the weights adjusted to reflect the relative importance of sensitivity and specificity for the researcher. Perkins and Schisterman (2006) proposed a weighted Youden's (1950) index based on disease prevalence and the costs of making a false positive or false negative; however, prevalence and misclassification costs are not always known. Another weighted Youden's (1950) index was proposed by D. Li et al. (2013) where the only additional information needed was to define the weights of sensitivity and specificity. The main concern with applying a weight is the justification of the weight chosen (D. Li et al., 2013).

TAP is defined as:

$$TAP = \frac{\left[1 - Y_{spec} - P\left(1 - Y_{spec} - Y_{sens}\right)\right]}{P}, \tag{2.3}$$

$$TAP = \frac{Y_{sens}}{PPV}, \tag{2.4}$$

where P is the disease prevalence in the population. The TAP index ranges from negative infinity to positive infinity with a score equal to one indicating the prevalence in the case definition equals the prevalence in the population (Campbell et al., 2008). The TAP index is unique as it incorporates the disease prevalence in the population (see equation 2.3); however, when the disease prevalence in the population is not known, TAP can be calculated using sensitivity and PPV (see equation 2.4). The TAP index focuses on selecting the case definition that produces the disease prevalence closes to the disease prevalence in the population (Campbell et al., 2008). This means that the case definition selected as optimal may not have the highest sensitivity or specificity but is the most accurate in identifying the disease prevalence (Campbell et al., 2008).

2.3 Regression Models for Meta-Analysis of Diagnostic Validation Studies

Meta-analysis is the process of combining results from independent studies to produce one overall estimate. In a meta-analysis of diagnostic validation studies, the estimates of sensitivity and specificity from studies of the same diagnostic test are combined to create a pooled estimate to evaluate the overall accuracy of the diagnostic test. Models that have been used to combine diagnostic validity estimates include: (a) univariate random-effects models, (b) bivariate random-effects models applied to sensitivity and specificity, and (c) trivariate random-effects models applied to sensitivity, specificity, and prevalence. The random-effects model accounts for dependence amongst the case definitions applied to the same population (i.e., repeated measurements on the case definitions), conditional on disease status; the model assumes that case definition characteristics (i.e., data source, length of the observation period for case ascertainment) are uncorrelated with the model parameters. Fixed-effects models are used when the studies have homogeneous characteristics (Menke, 2010).

A univariate model for the DOR produces a summary receiver operating characteristic (sROC) curve (Cleophas & Zwinderman, 2009; Reitsma et al., 2005), which plots the true positive rate against the false positive rate for multiple different thresholds of a diagnostic test. The ROC curve is typically used to determine the threshold with the best balance between the true positive and false positive rates. Typically, a linear model is fit to the log transform of the DOR. Cleophas and Zwinderman (2009) noted the benefits of modeling the DOR; correlation between sensitivity and specificity can be accounted for and covariates can be added to the model (Cleophas & Zwinderman, 2009).

Simulation has been used to examine the properties of separate univariate models applied to sensitivity and specificity estimates from a meta-analysis of diagnostic validity estimates

(Riley, Abrams, Sutton, Lambert, & Thompson, 2007). An empirical study has demonstrated the use of the univariate models and bivariate model applied to sensitivity and specificity to combine estimates across studies (Harbord et al., 2008). The logit transformation of sensitivity and specificity is assumed to be normally distributed (Harbord et al., 2008). The models produce separate mean estimates of sensitivity and specificity as well as their variances. Sensitivity and specificity are assumed to be randomly distributed, so that between-study heterogeneity can be accounted for in the model (Harbord et al., 2008).

A bivariate random-effects model simultaneously models sensitivity and specificity, which allows correlation between sensitivity and specificity to be accounted for. Two different distributions for sensitivity and specificity have been used: (a) normal distribution on the log transformed sensitivity and specificity (Menke, 2010) and (b) beta distribution (Kuss et al., 2014). Kuss et al. (2014) used simulation to compare two bivariate models: a) normal distribution with Pearson correlation and b) beta distribution with correlations calculated using copulas to describe the correlation structure between the marginal distributions; the authors tested three different copula models, including the Clayton, Gauss, and Placket copula. When correlation between sensitivity and specificity was present in the data, the models using the beta distribution with copula correlations tended to have lower bias and mean square error (MSE) of sensitivity and specificity than the normally distributed models (Kuss et al., 2014).

Simulation studies involving diagnostic validation studies have focused on estimation of five parameters: mean of sensitivity, mean of specificity, variance of sensitivity, variance of specificity, and correlation between sensitivity and specificity (Diaz, 2015; Kuss et al., 2014; Riley et al., 2007). These parameter estimates were used to compare the models on measures of bias, MSE, and 95% CI coverage of the mean. Mean bias in sensitivity and specificity is the

difference between the population mean and the estimated model mean. Kuss et al. (2014) considered bias to be large if it was above 3%. MSE of either sensitivity or specificity is calculated as the study variance divided by the number of simulations that converged. Coverage is approximated for sensitivity and specificity by assuming a t distribution (Riley et al., 2007); it is calculated as:

$$\hat{\beta}_j \mp \left(t_{n_j-1\,(0.975)} * \sqrt{var(\hat{\beta}_j)} \right), \tag{2.5}$$

where $\hat{\beta}_j$ is the estimate from the j^{th} simulation, $t_{n_j-1\,(0.975)}$ is the 97.5^{th} percentile on the *t* distribution with $n_j - 1$ degrees of freedom, and n_j is the number of studies in the j^{th} simulation. Lastly, the number of simulations that converged for each scenario were documented (Diaz, 2015; Kuss et al., 2014).

Harbord et al. (2008) compared univariate models of sensitivity and specificity to bivariate models using numeric examples. Harbord et al. (2008) applied the models to eight previously published meta-analyses and produced mean estimates of sensitivity and specificity with 95% CIs. The eight meta-analyses represented a variety of populations, health conditions, and used a different number of studies, ranged from 11 to 32 studies. The authors recommended the bivariate model over separate univariate models.

Riley et al. (2007) used simulation of sensitivity and specificity from a meta-analysis to demonstrate that the bivariate model had lower average bias, lower MSE, and higher 95% coverage compared to separate univariate models. The study also applied the models to numeric examples and found larger estimated variances of sensitivity and specificity for the bivariate models, which contradicted the simulation results (Riley et al., 2007). This study assumed logit transformed sensitivity and specificity to follow normal distributions and the models were

estimated using restricted maximum likelihood (REML) in SAS PROC MIXED (SAS Institute Inc., 2011).

The trivariate random-effects model is an extension of the bivariate random-effects models that includes prevalence as a third dependent variable. This model was developed in meta-analysis to account for changing disease prevalence between studies (Chu, Nie, Cole, & Poole, 2009; Hoyer & Kuss, 2015). Li and Fine (2011) showed that sensitivity and specificity are often correlated with disease prevalence, although this depends on the study population.

2.4 Rheumatoid Arthritis Diagnostic Validation Studies

Multiple studies have been conducted about the validity of RA diagnoses in administrative health data. These studies have been conducted in both adult and children/youth populations; they are summarized in Appendix A, Table A.1. Validation studies have been applied to diagnostic information in several administrative health databases including hospital discharge records, physician billing claims, pharmacy data, and laboratory data. These studies have assessed the accuracy of diagnoses recorded using the International Classification of Diseases 9th revision (ICD-9) or 10th revision (ICD-10). Specifically, these studies have identifying RA cases in administrative data using ICD-9 code 714 and ICD-10 codes M05 and M06.

Chart review was the most commonly used reference standard in the identified validation studies. The Canadian Community Health Survey was used as a reference standard in a one study (Lix et al., 2006); while another study from Nova Scotia used a clinical database to identify individuals with JRA (Stringer & Bernatsky, 2015). Lastly, one study used Bayesian latent class models to calculate sensitivity and specificity because a reference standard was not available (Bernatsky et al., 2014).

Lix et al. (2006) used Manitoba administrative health data to test the validity of 16 RA case definitions. Stringer (2015) tested just four case definitions to identify JRA in children less than 16 years of age in Nova Scotia; sensitivity estimates ranged from 0.53 to 0.86 and PPV estimates ranged from 0.52 to 0.65. Bernatsky et al. (2014) validated a single case definition for identifying RA patients in Quebec; the authors showed that sensitivity varied with the population characteristics of sex, age, and region of residence. Widdifield et al. (2013) tested 62 case definitions in Ontario administrative health data. Chart review from rheumatology clinics was the reference standard. This study identified the optimal case definition for individuals 20+ years of age as one hospitalization with an RA diagnosis code or three physician claims with at least one RA diagnosis code by a specialist in a 2-year period. This definition produced estimates for sensitivity, specificity, PPV, and NPV of 0.97, 0.85, 0.76, and 0.98, respectively. In 2014, Widdifield et al. conducted another study in which 48 case definitions were validated using Ontario administrative health data; the reference standard was chart review data from the Electronic Medical Record Administrative Linked Database (EMRALD). The optimal case definition was the same as identified in the previous study; however, this definition produced estimates of sensitivity, specificity, PPV, and NPV of 0.78, 1.00, 0.78, and 1.00, respectively.

2.5 Summary of Literature Review

Guidelines for reporting and evaluating diagnostic validation studies have been developed but provide little guidance about the number of case definitions that should be tested or how to select the optimal case definition when multiple case definitions are tested. Sensitivity and specificity are the measures most commonly reported in diagnostic validation studies; they are often used as the basis for selecting one case definition over another. Descriptive techniques

are primarily used to facilitate the selection process. Sensitivity and specificity are sometimes summarized into a single measure, including Youden's (1950) index and the DOR.

In meta-analyses of diagnostic validation studies, sensitivity and specificity have been modelled separately using univariate models and jointly using bivariate random-effects models. Simulation studies have demonstrated that bivariate random-effects models are less biased and more accurate than separate univariate models (Riley et al., 2007), but also are less likely to converge, making them more difficult to apply in practice.

Thirteen studies about the validity of RA diagnoses in administrative health data were identified from the literature; six of these were conducted using Canadian data and two studies focused on the accuracy of diagnoses in children and youth populations. The vast majority of these studies have tested the validity of multiple case definitions, using descriptive assessments of sensitivity and specificity to recommend one case definition.

CHAPTER 3 - METHODS

This research was conducted using computer simulation and a numeric example. In this section, both components of the research are described. The primary hypothesis is defined. Elements of the simulation study are described, including the statistical models, population parameters, model covariates, and statistical analysis. For the numeric example, we describe the data to be used, study design, and statistical analysis.

3.1 Hypothesis

My hypothesis in the simulation study was that the bivariate model of sensitivity and specificity would result in smaller bias and greater accuracy of the model parameter estimates than either the univariate models of sensitivity and specificity or the univariate model of Youden's (1950) summary measure of sensitivity and specificity.

3.2 Simulation Study

3.2.1 Statistical Models

The statistical models that were compared are: 1) univariate fixed-effects models independently applied to sensitivity and specificity, 2) bivariate mixed-effects model applied to sensitivity and specificity, and 3) univariate fixed-effects model applied to Youden's (1950) index, a summary measure of sensitivity and specificity. The univariate fixed-effects models for sensitivity and specificity was defined as:

$$Y_{l\,(sens)} = \beta_{0\,(sens)} + \beta_{m\,(sens)} X_{ml} + \varepsilon_{l\,(sens)}, \quad (3.1)$$

$$Y_{l\,(spec)} = \beta_{0\,(spec)} + \beta_{m\,(spec)} X_{ml} + \varepsilon_{l\,(spec)}, \quad (3.2)$$

where $Y_{l\,(sens)}$ and $Y_{l\,(spec)}$ are the values of sensitivity and specificity for the l^{th} case definition, $\beta_{0\,(sens)}$ is the marginal intercept for sensitivity, $\beta_{m\,(sens)}$ ($m = 1,\ldots,k$) is the marginal parameter for the m^{th} explanatory variable for sensitivity, $\beta_{0\,(spec)}$ is the marginal intercept for specificity,

$\beta_{m\,(spec)}$ is the marginal parameter for the m^{th} explanatory variable for specificity, X_{ml} is the value of the explanatory variable for the m^{th} covariate and the l^{th} case definition, and $\varepsilon_{l\,(sens)}$ and $\varepsilon_{l\,(spec)}$ are the model error terms for sensitivity and specificity, respectively.

The bivariate mixed-effects model was:

$$Y_{il} = \beta_{0i} + u_{il} + \beta_{mi}X_{ml} + \varepsilon_{il}, \tag{3.3}$$

where Y_{il} is the value of sensitivity when i=1 and specificity when i=2 for the l^{th} case definition, β_{0i} is the marginal intercept for sensitivity when i=1 and specificity when i=2, u_{il} is the random-effect estimate of sensitivity when i=1 and specificity when i=2 for the l^{th} case definition, β_{mi} ($m = 1,\ldots,k$) is the marginal parameter for the m^{th} explanatory variable for sensitivity when i=1 and specificity when i=2, X_{ml} is the value of the explanatory variable for the m^{th} predictors and the l^{th} case definition, and ε_{il} is the error term for sensitivity when i=1 and specificity when i=2.

The univariate fixed-effects model for Youden's (1950) index was:

$$W_l = \beta_{0(W)} + \beta_{m(W)}X_{ml} + \varepsilon_{l(W)}, \tag{3.4}$$

where W_l is the value of Youden's (1950) index for the l^{th} case definition, $\beta_{0(W)}$ is the marginal intercept for Youden's (1950) index, $\beta_{m\,(W)}$ ($m = 1,\ldots,k$) is the marginal parameter for the m^{th} explanatory variable for Youden's (1950) index, X_{ml} is the value of the explanatory variable for the m^{th} predictors and the l^{th} case definition, and $\varepsilon_{l(W)}$ is the error term for Youden's (1950) index.

3.2.2 Simulating Values of Sensitivity and Specificity

The models were applied to simulated values of sensitivity and specificity from correlated beta distributions. The beta distribution is defined as:

$$f(y;\alpha,\gamma) = \frac{1}{B(\alpha,\gamma)}y^{\alpha-1}(1-y)^{\gamma-1}, \tag{3.5}$$

where $f(y;\alpha,\gamma)$ is the probability density function for y ($0 \leq y \leq 1$) and shape parameters α and γ, and $B(\alpha,\gamma)$ is the beta function constant. This distribution has been used in recent studies that have modelled sensitivity and specificity (Hoyer & Kuss, 2015; Kuss et al., 2014; J. Li & Fine, 2011). This study adopted the methodology described by Wicklin (2013). The first step was to simulate values from standard normal distributions, Z_1 and Z_2. Define the correlation between Z_1 and Z_2 as ρ. Then:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim Normal\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right). \tag{3.6}$$

The second step was to apply the following fundamental probability theorem: If F is the cumulative distribution function of a continuous random variable, Z, then the random variable U = F(Z) is distributed as U(0,1) (Weiss, Holmes, & Hardy, 2005; p471). The theorem was used to transform the bivariate normal distribution to the bivariate uniform distribution via the cumulative distribution function (CDF).

In the last step, the inverse CDF was used to transform the correlated uniform distributions to the desired beta distributions for Y_{sens} and Y_{spec}, with specified shape parameters, $\alpha_{sens}, \gamma_{sens}, \alpha_{spec}$, and γ_{spec}, from the population mean and variance of sensitivity and specificity. The distribution of sensitivity and specificity was:

$$\begin{pmatrix} Y_{sens} \\ Y_{spec} \end{pmatrix} \sim Beta\left(\begin{pmatrix} \alpha_{sens} \\ \alpha_{spec} \end{pmatrix}, \begin{pmatrix} \gamma_{sens} \\ \gamma_{spec} \end{pmatrix}\right). \tag{3.7}$$

The beta distributions were correlated because sensitivity and specificity have shown to be dependent in practice. A second approach to model sensitivity and specificity is to use a bivariate beta distribution; however, little, if any literature has explored the properties of this distribution for modeling sensitivity and specificity.

The shape parameters for the beta distributions were calculated from the population means and variances of sensitivity and specificity specified for each simulation condition:

$$\alpha_{sens} = \left(\frac{1-\mu_{sens}}{\sigma^2_{sens}} - \frac{1}{\mu_{sens}}\right)\mu^2_{sens}, \tag{3.8}$$

$$\gamma_{sens} = \left(\frac{1}{\mu_{sens}} - 1\right)\alpha_{sens}, \tag{3.9}$$

where α_{sens} and γ_{sens} are shape parameters for the beta distribution of sensitivity, μ_{sens} is the mean of sensitivity, and σ^2_{sens} is the variance of sensitivity. Then:

$$\alpha_{spec} = \left(\frac{1-\mu_{spec}}{\sigma^2_{spec}} - \frac{1}{\mu_{spec}}\right)\mu^2_{spec}, \tag{3.10}$$

$$\gamma_{spec} = \left(\frac{1}{\mu_{spec}} - 1\right)\alpha_{spec}, \tag{3.11}$$

where α_{spec} and γ_{spec} are the shape parameters for the beta distribution of sensitivity, μ_{spec} is the mean of specificity, and σ^2_{spec} is the variance of specificity.

Youden's (1950) index was calculated from the simulated values of sensitivity and specificity. Youden's (1950) index follows a beta distribution:

$$W \sim Beta\ (\alpha_W, \gamma_W), \tag{3.12}$$

where α_W and γ_W are the shape parameters calculated from the population mean and variance of Youden's (1950) index. When sensitivity and specificity were correlated, the population mean and variance of Youden's (1950) index were defined as (Weiss et al., 2005; p355):

$$\mu_W = \mu_{sens} + \mu_{spec} - 1, \tag{3.13}$$

$$\sigma^2_W = E(W^2) - E(W)^2$$

$$= E\left[\left(Y_{sens} + Y_{spec} - 1\right)^2\right] - \left[E\left(Y_{sens} + Y_{spec} - 1\right)\right]^2$$

$$= E(Y^2_{sens}) + E\left(Y^2_{spec}\right) - 2E(Y_{sens}) - 2E\left(Y_{spec}\right) + 2E\left(Y_{sens}Y_{spec}\right) + 1 - \mu^2_W, \tag{3.14}$$

where μ_W, μ_{sens}, and μ_{spec} are the population means of Youden's (1950) index, sensitivity, and specificity, respectively. When sensitivity and specificity were independent, the mean and variance of Youden's (1950) index were (Weiss et al., 2005; p355):

$$\mu_W = \mu_{sens} + \mu_{spec} - 1, \quad (3.15)$$

$$\sigma_W^2 = \sigma_{sens}^2 + \sigma_{spec}^2, \quad (3.16)$$

where σ_W^2, σ_{sens}^2, and σ_{spec}^2 are the population variances of Youden's (1950) index, sensitivity, and specificity, respectively. When sensitivity and specificity were independent, the variance of Youden's (1950) index (equation 3.16) was a special case of the variance when sensitivity and specificity were correlated (equation 3.14).

Youden's (1950) index can take on a value in the range [-1, +1] (Chen et al., 2015); however, the majority of studies indicate values will range from 0 to 1 (Lai, Tian, & Schisterman, 2012; Schisterman & Perkins, 2007; Youden, 1950). Chen et al. (2015) notes that even though values of Youden's (1950) index can fall below zero, a value in the range from -1 to 0 has no interpretation. A negative value arises when both sensitivity and specificity are below 0.5, or when either sensitivity or specificity is very small; both situations arise infrequently. When these situations occur, no interpretation is given as it is associated with poor sensitivity and or specificity and the case definition would not be considered for use in the administrative health data. In the simulation study, a negative value for Youden's (1950) index was set to zero. Values of zero were given since negative values could not be included in the beta distribution of Youden's (1950) index and a negative value has the same interpretation as a value of zero.

3.2.3 Simulation Parameter Values

The following parameters were investigated in the simulation study: mean and variance of sensitivity, mean and variance of specificity, correlation between sensitivity and specificity, sample size (i.e., number of case definitions), number of covariates, and distributional properties of the covariates. Each parameter and its selected values are shown in Table 3.1.

Table 3.1: Simulation study parameters and values

Parameter	Values
μ_{sens}	0.70, 0.80, 0.90
μ_{spec}	0.90
σ^2_{sens}	0.01, 0.03
σ^2_{spec}	0.01, 0.03
$\rho_{sens\ spec}$	0, -0.20, -0.70
N	40, 75
μ_{X_1}	0.40
μ_{X_2}	0.60
μ_{X_3}	0.70
$\rho_{X_1X_2}$	0.20
$\rho_{X_1X_3}$	0.05
$\rho_{X_2X_3}$	-0.20

Sensitivity means were set to 0.70, 0.80, and 0.90 with variances of 0.01 and 0.03. Specificity means were set to 0.90 with variances of 0.01 and 0.03. Three values of the sensitivity-specificity correlation were considered: 0, -0.20, and -0.70. The sample size (i.e., number of case definitions) was set at 40 and 75.

Each statistical model included up to three covariates, X_1, X_2, and X_3, which were simulated using the following distributions:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim Binary \begin{pmatrix} 0.40 \\ 0.60 \\ 0.70 \end{pmatrix}. \tag{3.17}$$

We defined models with: 1) X_1 and X_2. 2) X_2 and X_3, and 3) X_1, X_2, and X_3, as well as the null model (i.e., intercept only). Increasing the number of covariates can affect the precision of the estimates and/or model convergence properties (Montgomery, 2012). Only binary covariates were included in the models, because covariates for describing case definition characteristics are commonly binary in nature.

A SAS program called RandMVBinary was used to simulate the correlated binary variables (Emrich & Piedmonte, 1991; Wicklin, 2013). This program was developed based on research by Emrich & Piedmonte (1991) to generate correlated binary variables and was written by Rick Wicklin (2013). This program requires a vector of the parameter probabilities and a matrix of the parameter correlations to be defined and simulates the distributions using a matrix of zeros and ones. This process is completed in a similar manner as described previously for correlating the beta distributions using transformations (see Appendix C).

Overall, 36 scenarios were considered for each specification of the model (i.e., null, two covariates and three covariates), as shown in Appendix A, Table A.2. Accordingly a total of 144 scenarios were investigated for each of the three statistical models. There were 2,000 replications conducted for each scenario.

3.2.4 Selection of Simulation Parameter Values

The parameter values for the simulation study were chosen based on a review of RA validation studies. Seven of the 11 (adult) RA diagnostic validation studies identified from the literature review reported sensitivity and specificity estimates for the investigated case definitions; a total of 143 case definitions were tested in these studies. The means values of sensitivity and specificity selected for the simulation were based on frequent combinations of sensitivity and specificity (Table 3.2).

To assess how sensitivity and specificity were related, their variances and correlations were calculated from each study (Table 3.3). The variances ranged from 0.00 to 0.15, with most values between 0.01 and 0.04. Accordingly, small and large variances were chosen for sensitivity and specificity in the simulation study. The correlation between sensitivity and specificity was

always negative, and ranged from -0.06 to -0.93. Given this wide range of values, a small and large negative correlation was chosen for the simulation study.

Table 3.2: Frequency of sensitivity and specificity combinations for n = 143 case definitions tested in RA diagnostic validation studies

$Y_{l\,(sens)}$	$Y_{l\,(spec)}$		
	0.70-0.79	0.80-0.89	0.90-1.00
0.00-0.29	0	0	**29**
0.30-0.49	0	0	3
0.50-0.69	0	1	6
0.70-0.79	2	3	**20**
0.80-0.89	2	1	**21**
0.90-1.00	7	**32**	**16**

Table 3.3: Mean, variance, and estimated correlation of sensitivity and specificity in RA diagnostic validation studies

Study	N	$\hat{\mu}_{sens}\,(\hat{\sigma}^2_{sens})$	$\hat{\mu}_{spec}\,(\hat{\sigma}^2_{spec})$	$\rho_{sens\,spec}$
Gabriel, 1994	1	0.89 (0.00)	0.74 (0.00)	-
Lix, 2006	16	0.08 (0.00)	0.99 (0.00)	-0.70
Singh, 2004	5	0.85 (0.01)	0.84 (0.03)	-0.93
Ng, 2012	5	0.52 (0.04)	0.85 (0.02)	-0.06
Widdifield, 2013	61	0.91 (0.02)	0.85 (0.01)	-0.32
Widdifield, 2014	43	0.75 (0.02)	0.99 (0.00)	-0.33
Carrara, 2015	18	0.43 (0.15)	0.92 (0.04)	-0.47

Features of the case definitions included as model covariates were: type of data source, number of contacts, specialist contacts, duration of time between contacts, and length of observation period. The case definition parameters from the seven RA validation studies that recorded both sensitivity and specificity were further explored to examine their frequency and probability of being used (Table 3.4).

Correlations between the binary covariates were based on the correlations estimated from the studies by Widdifield et al. (2014; 2013) (Table 3.5). These two studies were chosen as the basis for the simulations because they are the only studies to have tested more than 30 RA

definitions each. Accordingly, these studies provide more accurate estimates of correlation values than studies that tested a smaller number of case definitions. The correlations ranged from -0.39 to 0.36; therefore, both negative and positive correlations were selected for the simulation study.

Table 3.4: Frequency of case definition characteristics in RA diagnostic validation studies

Study	Hospital data (probability)	Physician data (probability)	Pharmacy data (probability)	Specialist (probability)	Visit separation (probability)
Gabriel, 1994	1 (1.00)	1 (1.00)	0 (0.00)	0 (0.00)	0 (0.00)
Singh, 2004	4 (0.80)	4 (0.80)	3 (0.60)	0 (0.00)	0 (0.00)
Lix, 2006	4 (0.25)	16 (1.00)	4 (0.25)	0 (0.00)	0 (0.00)
Ng, 2012	5 (1.00)	5 (1.00)	3 (0.60)	4 (0.80)	5 (1.00)
Widdifield, 2013	38 (0.62)	59 (0.97)	30 (0.49)	28 (0.46)	8 (0.13)
Widdifield, 2014	26 (0.60)	41 (0.95)	15 (0.35)	19 (0.44)	7 (0.16)
Carrara, 2015	7 (0.39)	9 (0.50)	14 (0.78)	8 (0.44)	0 (0.00)

Table 3.5: Correlations between binary case definition parameters for RA diagnostic validation studies

a) Widdifield, Bombardier, Bernatsky et al., 2014

	Hospital	Physician	Pharmacy	Specialist	Separation
Hospital	1.0000	-0.1786	0.2924	0.0490	0.2277
Physician	-0.1786	1.0000	0.1617	0.1965	0.0974
Pharmacy	0.2924	0.1617	1.0000	-0.0617	-0.1906
Specialist	0.0490	0.1965	-0.0617	1.0000	-0.3924
Separation	0.2277	0.0974	-0.1906	-0.3924	1.0000

b) Widdifield, Bernatsky, Paterson et al., 2013

	Hospital	Physician	Pharmacy	Specialist	Separation
Hospital	1.0000	-0.1432	0.3594	0.0378	0.1019
Physician	-0.1432	1.0000	0.1811	0.1696	0.0715
Pharmacy	0.3594	0.1811	1.0000	-0.1165	-0.1879
Specialist	0.0378	0.1696	-0.1165	1.0000	-0.3579
Separation	0.1019	0.0715	-0.1879	-0.3579	1.0000

3.2.5 Model Fitting

SAS IML version 9.3 (SAS Institute Inc., 2011) was used to simulate the data. All analyses were completed with SAS PROC GLIMMIX using the quadrature iterative method for parameter estimation, as recommended by Zhang et al. (2011). The quadrature estimation method uses maximum likelihood estimation to approximate an integral over a fixed or varying number of quadrature points. The quadrature method is based on the log-likelihood approximation method, which results in estimates for common log-likelihood fit statistics such as the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) (SAS Institute Inc., 2011; Zhang et al., 2011). The number of quadrature points is flexible and can be increased to improve the quality of the approximation to the log-likelihood function; however, increasing the number of quadrature points or having many random effects makes the model computation more difficult and less likely to produce reliable estimates of the random effects (Diaz, 2015; Zhang et al., 2011).

The mixed-effects model used the Cholesky decomposition to ensure that the estimated variance-covariance matrix of the random effects was positive semi-definite and the model converged (Menke, 2010). The SAS/STAT 9.3 User's Guide (2011) recommends the Cholesky variance-covariance structure. The Cholesky structure has been used in previous meta-analysis of random effects models for diagnostic validation studies (Menke, 2010; Riley et al., 2007).

3.2.6 Model Evaluation

For each simulation scenario, the models were compared on the following measures of performance: percent of simulation replications that converged, bias, mean square error (MSE), and 95% CI coverage. The percent of simulation replications that converged was calculated as

the number of models that converged out of the 2,000 replications. Bias, MSE, and 95% CI coverage were calculated for each of the model parameters. Bias was calculated as:

$$Bias = (\bar{\hat{\beta}} - \beta) * 100, \tag{3.18}$$

where $\bar{\hat{\beta}}$ is mean estimate of bias based on the 2,000 simulation replications and β is the parameter value. MSE was calculated as:

$$MSE = [(\bar{\hat{\beta}} - \beta)^2 + (SE(\hat{\beta}))^2] * 100, \tag{3.19}$$

where $SE(\hat{\beta})$ is the standard error of the parameter estimate. The Delta method was used to calculate the standard errors in PROC GLIMMIX (SAS Institute Inc., 2011). The 95% CI coverage of the mean was calculated as the percentage of times that the true mean value was contained in the 95% CI. The 95% CI was calculated as

$$95\%\ CI = \bar{\hat{\beta}} \mp t_{(n-1)1-\alpha/2} SE(\hat{\beta}), \tag{3.20}$$

where $t_{(n-1)1-\alpha/2}$ is the critical value from the t-distribution at the $1 - \alpha/2$ percentile with $n - 1$ degrees of freedom and α is the level of significance.

3.2.7 Modeling of the Simulation Results

Descriptive analyses of the simulation results were conducted using means of the model evaluation statistics for each of the simulation parameters (sample size, sensitivity mean, variance of sensitivity and specificity, and correlation between sensitivity and specificity). Additionally, the results were graphically described.

Analysis of variance (ANOVA) models were applied to the measures of model performance. The ANOVA models were used to test the association of the simulation parameters (i.e., sample size, sensitivity mean, sensitivity and specificity variance, and correlation between sensitivity and specificity) with model performance. Thirteen ANOVA models were fit to the data for each of the five model dependent variables: univariate sensitivity, univariate specificity,

Youden's (1950) index, sensitivity in the bivariate model, and specificity in the bivariate model. The outcome measures in each model were the model performance statistics: percent of model convergence, intercept bias, intercept MSE, intercept 95% CI coverage, X_1 coefficient bias, X_1 coefficient MSE, X_1 coefficient 95% CI coverage, X_2 coefficient bias, X_2 coefficient MSE, X_2 coefficient 95% CI coverage, X_3 coefficient bias, X_3 coefficient MSE, and X_3 coefficient 95% CI coverage.

The ANOVA model covariates included the simulation parameters: sample size, mean of sensitivity, variance of sensitivity and specificity, and correlation between sensitivity and specificity. Specificity mean was not used as a covariate as it remained constant at 0.90 throughout the simulation scenarios. The variance of sensitivity and specificity were combined into one covariate because the values were always equal throughout the simulation scenarios.

The ANOVA models were fit using forward selection techniques to identify the number of covariate interactions to include in the model (Coombs & Algina, 1996; Leite et al., 2015; Shi, Leite, & Algina, 2010). The forward selection method was applied in three steps: 1) all main effects were included in the model, 2) all main effects and two-way interactions were included in the model, and 3) all main effects, two-way interactions, and three-way interactions were included in the model. The optimal model was chosen based on the percent of explained variation (i.e., R^2) value with consideration of overfitting the model with the interaction terms. When two- and three-way interactions were included in the model with four covariates and five covariates, 14 and 25 parameters had to be estimated, respectively. Inclusion of the three-way interactions may overfit the ANOVA model.

The ANOVA models for intercept bias, intercept MSE, intercept 95% CI coverage were fit to include an additional dichotomous (i.e., binary) variable indicating if at least one covariate

of X_1, X_2, or X_3 was used in the simulation scenario. This covariate was included as it indicated the change in the simulation design and it accounted for the percent explained variation from the addition of covariates X_1, X_2, or X_3 in the model.

Once the optimal model(s) were selected, the percent explained variance and p-value of each parameter estimate was recorded. Percent explained variance was calculated by the parameter estimates Type 3 sum of squares divided by the total sum of squares and multiplied by 100. Type 3 sum of squares were used because they test each factor (main effects and interactions) while conditioning on all other factors in the model (i.e., holding all other factors constant) (SAS Institute Inc., 2011). Type 3 sum of squares is most commonly used when interactions between covariates are of interest and are statistically significant. Parameters that had $p \leq 0.0001$ were considered to be statistically significant. This strict criterion was chosen because many tests were conducted and it limits the chance of a false positive (Shaffer, 1995). When a simulation parameter was statistically significant and had a percent explained variance > 1%, the simulation parameter was considered as a factor that affected the model evaluation statistic (i.e., bias, MSE, 95% CI coverage, and model convergence).

3.3 Numeric Example

The numeric example methods are described in three sections: data sources, study variables, and statistical analysis.

3.3.1 Data Sources

The three models selected for this research were applied to a secondary analysis of case definitions from a RA diagnostic validation study (Widdifield, Bernatsky, et al., 2013). This validation study was conducted using Ontario administrative health data from April 1, 1991 to March 31, 2011. Physician billing claims, hospital discharge abstracts, and emergency room

(ER) records were available for all individuals in the study and pharmacy data was also available for individuals 65+ years. The reference standard data source was medical records. A total of 450 medical records were reviewed from 18 rheumatology clinics; 149 individuals had a confirmed RA diagnosis.

3.3.2 Study Variables

The published paper reported results for 61 case definitions. The first author was contacted to obtain additional case definitions not reported in the publication. This request resulted in 87 additional case definitions for the secondary analysis.

In total, 148 case definitions were tested; 57 of these definitions were tested for individuals 20+ years and the remaining 91 case definitions were defined for individuals 65+ years. All case definitions tested in the 20+ years age group were tested again in the 65+ years age group. The additional 34 case definitions included prescription medication characteristics, such as whether the individual filled a prescription for a steroid, disease-modifying antirheumatic drug (DMARD) medication, or biological agent. These case definitions were not tested in the 20+ years age group because the prescription medication data were only available for individuals 65+ years. The case definition characteristics and their values used in the secondary analysis are shown in Table 3.6.

A dataset was constructed that included coded case definition characteristics, and estimates of sensitivity, specificity, and Youden's (1950) index for each case definition. The age group for each case definition (20+ or 65+) was also coded. Sensitivity, specificity, and Youden's (1950) index were used as dependent variables in the univariate and bivariate models. The 12 case definition characteristics and the age group binary variable were used as model covariates.

Table 3.6: Case definition characteristics and their values in the secondary analysis

Characteristic	Values
Number of hospital discharge records	0, 1+
Number of diagnoses in ER records	0, 1+
Number of diagnoses in physician claims	0, 1+
Number of diagnoses in physician claims	0, 1+, 2+, 3+
Length of observation period to identify the physician diagnoses	0, 1 year, 2 years, 3 to 5 years, ever
Number of specialist diagnoses	0, 1+
Number of specialist diagnoses	0, 1+, 2+, 3+
If a time separation between two physician claims was required	Yes, no
Use of exclusion criteria A	Yes, no
Use of exclusion criteria B	Yes, no
Number of prescriptions for a steroid, DMARD, or biological agent	0, 1+
Number of prescriptions for DMARD or biological agent (not including steroid prescriptions)	0, 1+

3.3.3 Statistical Analysis

Descriptive analyses of the case definition characteristics were conducted using frequencies, percentages, means, and correlations. Spearman correlations were calculated for the case definition characteristics.

The univariate models of sensitivity and specificity were fit to the data as per equations 3.1 and 3.2. The bivariate mixed-effects model was fit to the data as per equation 3.3. Youden's (1950) index was calculated and a univariate fixed-effects model was fit to the data as per equation 3.4. The dependent variables were fit to follow a beta distribution with a logit link function. Thus, all model parameter estimates were reported using the logit scale.

First, null (i.e., intercept-only) models were fit to the data. Then, the covariates were added to each model separately. Then, multivariable models were fit using the covariates with $p \leq 0.01$ based on the previous models. Covariates in the multivariable (i.e., adjusted) models that

had $p \leq 0.0001$ were judged to be statistically significant. This strict criterion was chosen to limit the probability of a Type I error (Shaffer, 1995).

Model fit was assessed for using the AIC (Akaike, 1974), BIC (Schwarz, 1978), and likelihood ratio test (LRT); smaller values of the AIC and BIC and a statistically significant likelihood ratio statistic indicate a better fitting model. The LRT is sensitive to sample size, which is why it was used in combination with the AIC and BIC. The bivariate and univariate models were estimated using the quadrature iterative estimation method and variance-covariance structure used in the simulation study. All analyses were conducted using SAS version 9.3 (SAS Institute Inc., 2011).

3.4 Ethical Considerations

The RA example used results from published validation studies. The sensitivity and specificity values used from the case definitions are publically available. These values cannot be used to identify any individuals. Ethical approval was not required for this study.

CHAPTER 4 – RESULTS

4.1 Simulation Study

Results for the simulation study are described in two sections, the first section focuses on the descriptive analyses and the second focuses on the statistical models. The descriptive analysis section describes the relationships between the model evaluation statistics and the simulation parameters for each model (Figure 4.1 to Figure 4.4). The statistical modeling section formally tests these relationships using ANOVA models (Table 4.1 to Table 4.6).

4.1.1 Descriptive Analyses

This section describes each statistical model by average bias in the intercept, average MSE in the intercept, average 95% CI coverage in the intercept, and average model convergence stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation. Additionally, figures showing the average variance for the intercept bias and the average variance for the intercept MSE are included in the Appendix B Figures B.1 and B.2, respectively.

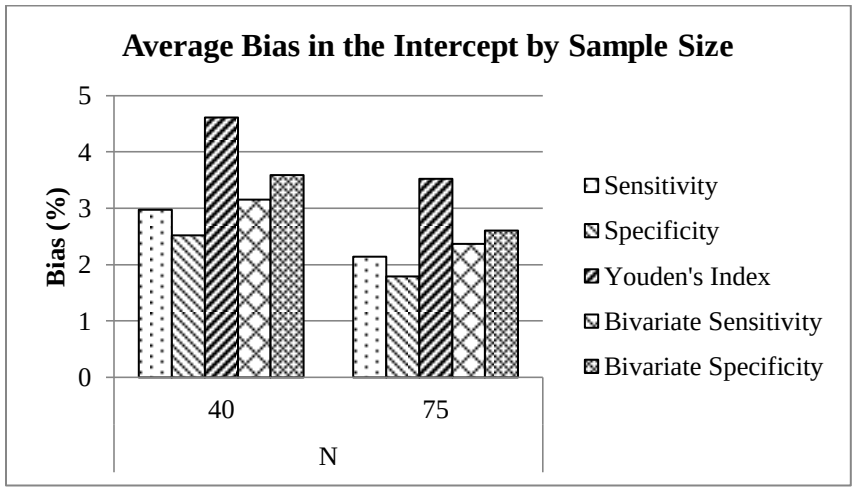
Similar descriptive statistics were produced for covariates X_1, X_2, and X_3 (see Appendix B). Appendix B Figure B.3 to Figure B.5 describe each statistical model by average bias, average MSE, and average 95% CI coverage in X_1 stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation. Appendix B Figure B.6 to Figure B.8 describe each statistical model by average bias, average MSE, and average 95% CI coverage in X_2 stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation. Appendix B Figure B.9 to Figure B.11 describe each statistical model by average bias, average MSE, and average 95% CI coverage in X_3 stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation.

In the simulation study, a population sensitivity mean of 0.50 was also tested. These scenarios were not reported as they showed similar results and provided no additional knowledge to the simulation study.

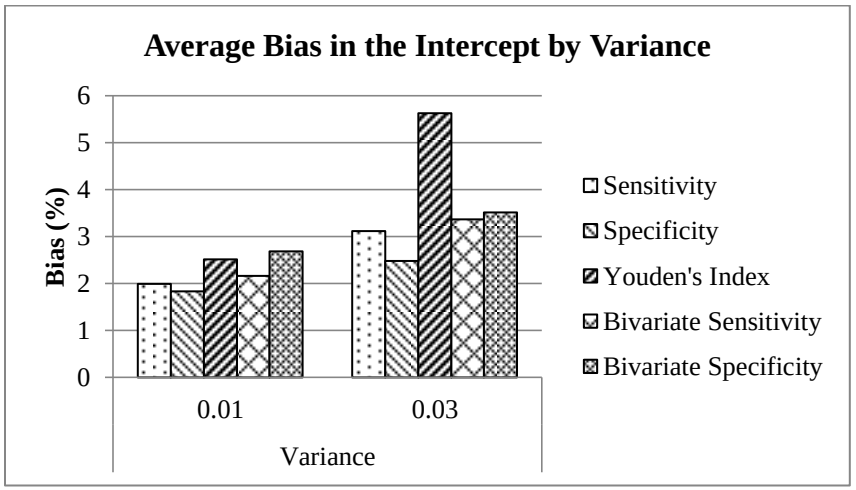
The bias of the intercept was largest for the univariate model of Youden's (1950) index and the smallest for specificity in the bivariate model (Figure 4.1). Average bias remained small over all statistical models and all simulation parameters. Change in the simulation parameter values produced only a small change in the average bias.

Figure 4.1: Average bias in the intercept for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

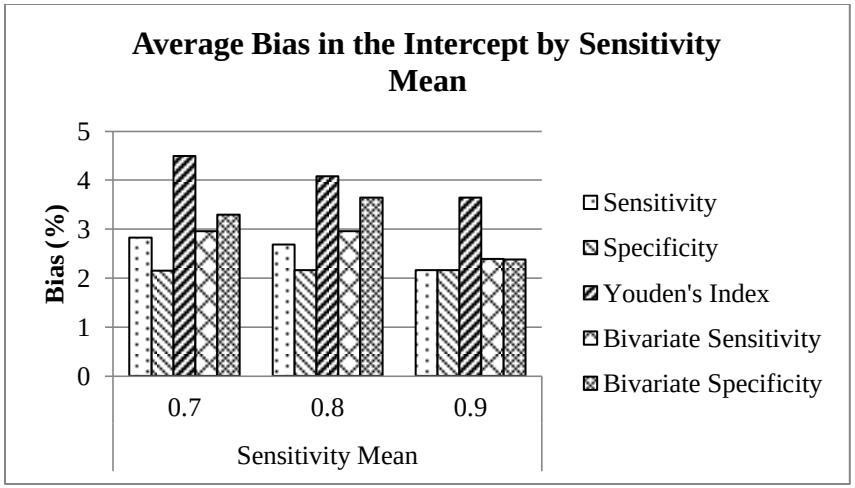
a)



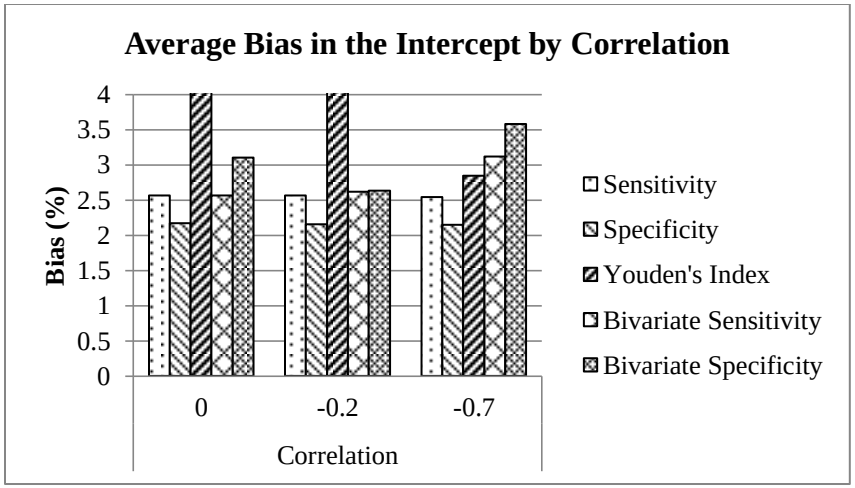
b)



c)



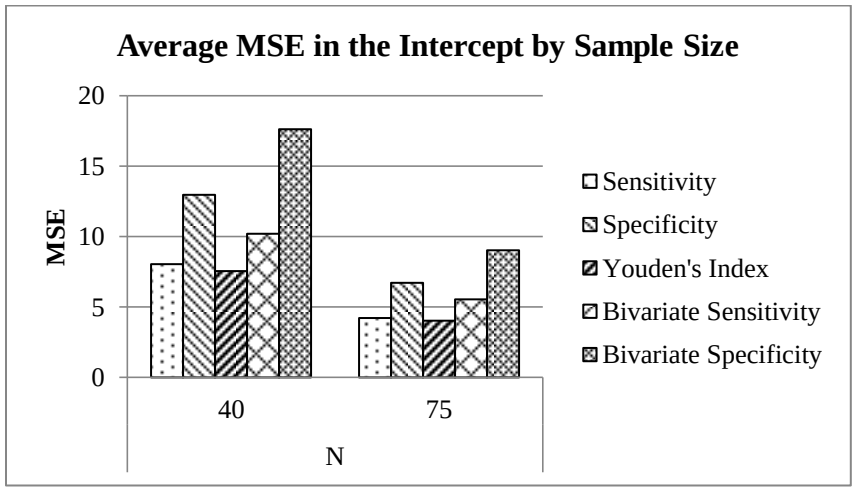
d)



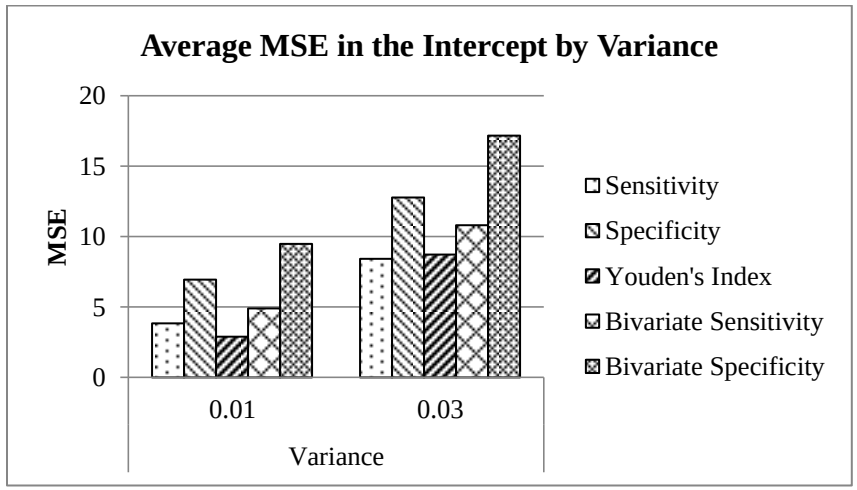
The MSE of the intercept was largest for specificity in the bivariate model and smallest for Youden's (1950) index (Figure 4.2). The simulation parameters showed little change in MSE of the intercept on the univariate models and sensitivity in the bivariate model. MSE of the intercept of specificity in the bivariate model decreased as each of the simulation parameters increased.

Figure 4.2: Average mean square error (MSE) of the intercept for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

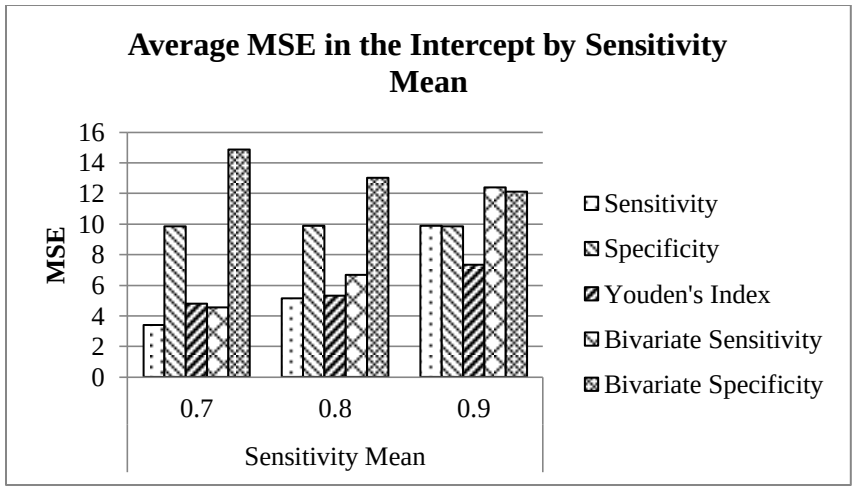
a)



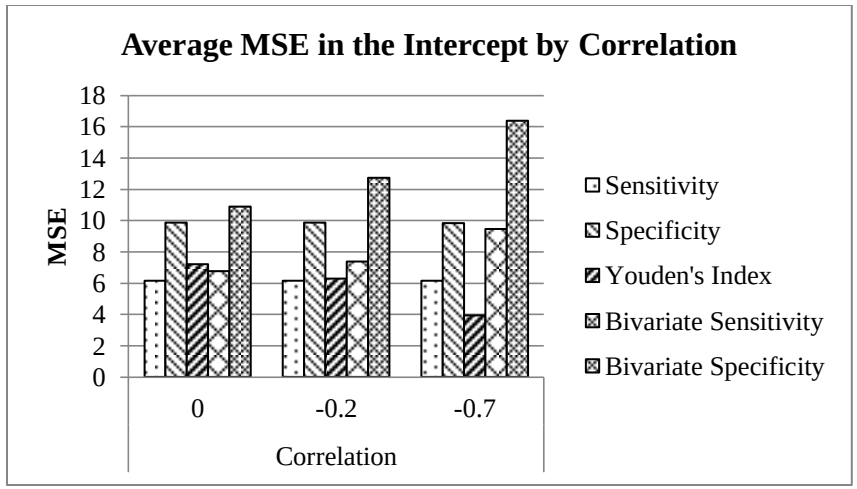
b)



c)



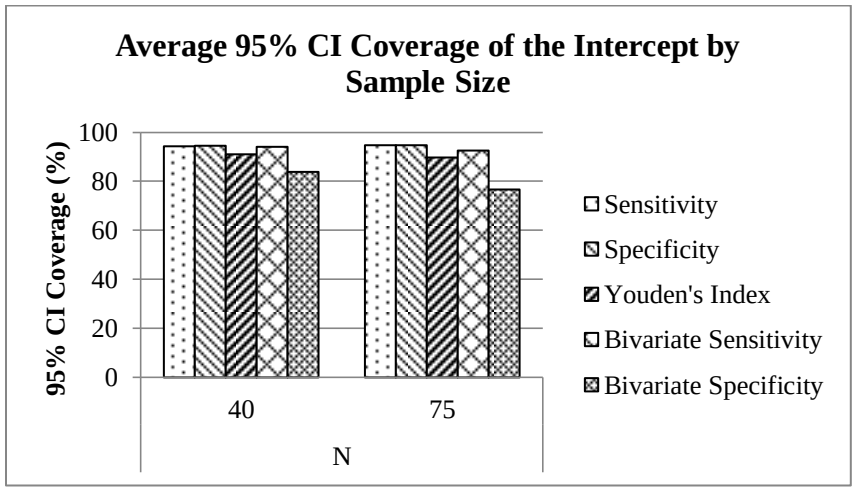
d)



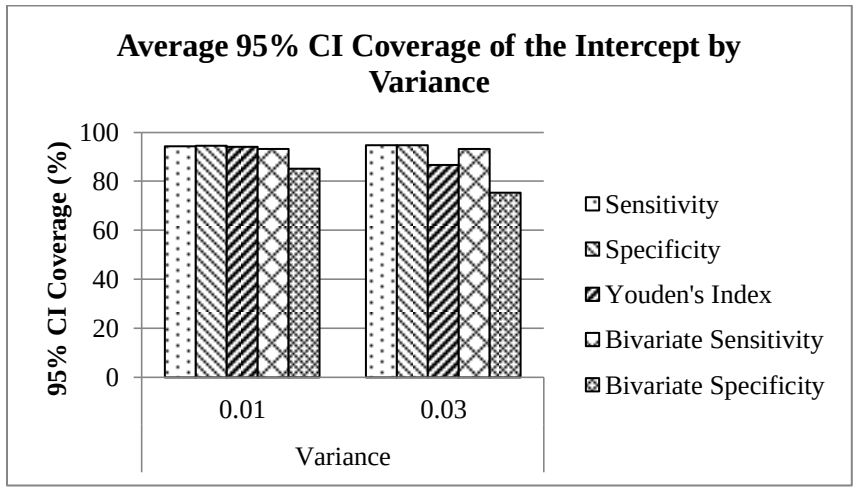
The univariate models and sensitivity in the bivariate model showed similar average 95% CI coverage for the intercept (Figure 4.3). The average 95% CI coverage for the intercept of specificity in the bivariate model was always less than 90.0%. The average 95% CI coverage for the intercept of specificity in the bivariate model decreased when the variance increased and the mean of sensitivity decreased.

Figure 4.3: Average 95% confidence interval (CI) coverage of the intercept for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

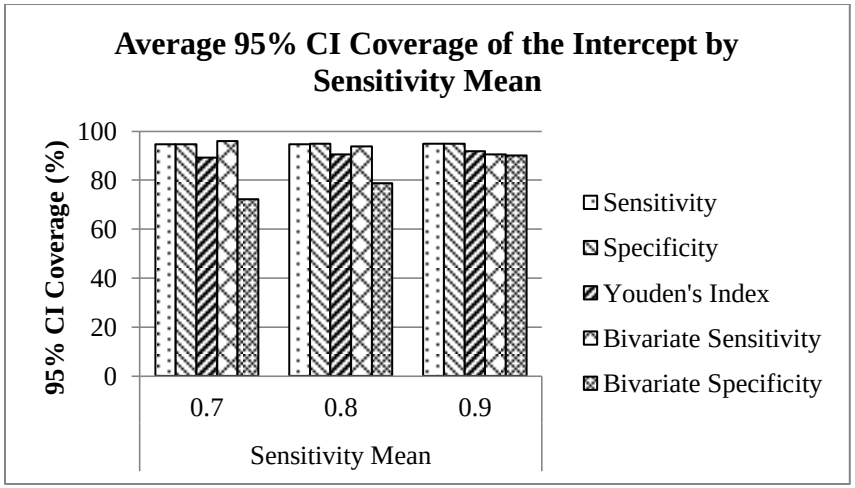
a)



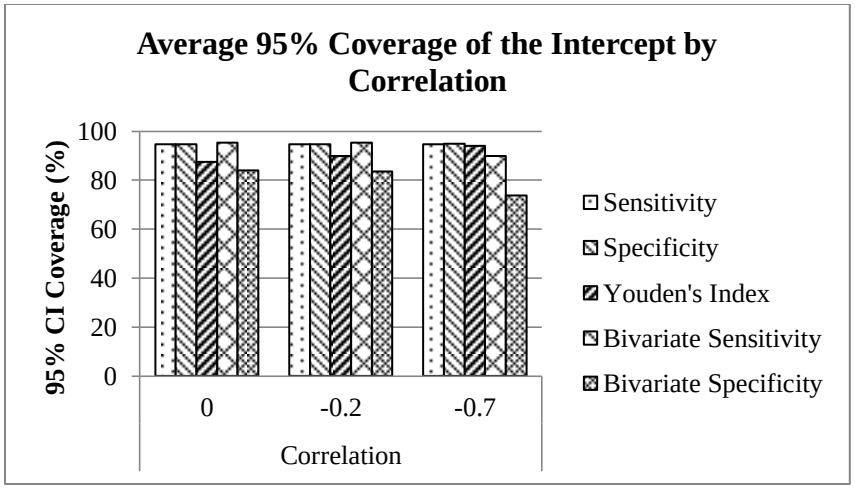
b)



c)

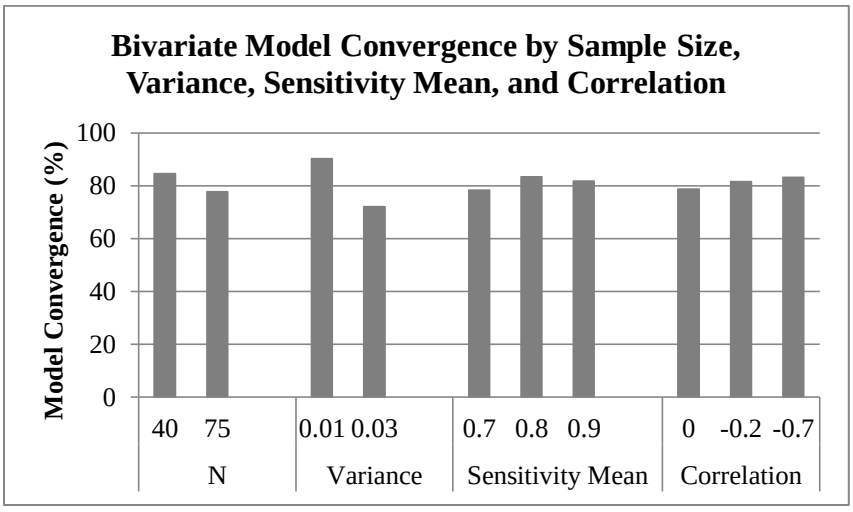


d)



The univariate models of sensitivity, specificity, and Youden's (1950) index converged 100% of the time. However, the bivariate models did not always converge; a minimum of 25% and a maximum of 99% of the replications converged. Figure 4.4 shows that the convergence percent decreased as sample size and variance increased. The convergence percent increased slightly as the sensitivity mean and correlation increased.

Figure 4.4: Average percentages of bivariate model convergence stratified by sample size, variance, sensitivity mean, and correlation



4.1.2 Statistical Modeling

This section describes the ANOVA models that were conducted to further explore the relationships between the simulation evaluation statistics and the simulation parameters. The first step was to select interactions (none, 2-way, or 3-way) to include in the models; this decision was informed by the percent explained variation. R^2 increased slightly as levels of interactions were added for the intercept bias, intercept MSE and intercept 95% CI coverage in each of the models (Table 4.1). All ANOVA models including main effects 2-way interactions were chosen because adding the three-way interaction showed little gain in R^2.

Table 4.1: Percent explained variation (R^2) in bias, mean squared error (MSE), and 95% confidence interval (CI) coverage by ANOVA models with main effects, two-way interaction, and three-way interactions for the intercept

Model	**Univariate Sensitivity R^2 (Δ %)**	**Univariate Specificity R^2 (Δ %)**	**Univariate Youden's Index R^2 (Δ %)**	**Bivariate Sensitivity R^2 (Δ %)**	**Bivariate Specificity R^2 (Δ %)**
			Bias		
Main effects	77.92	82.23	84.62	73.95	15.93
+ 2-way interaction	85.22 (7.30)	84.77 (2.54)	93.08 (8.46)	85.24 (11.29)	26.94 (11.01)
+ 3-way interaction	86.28 (1.06)	84.81 (0.04)	93.82 (0.74)	86.70 (1.46)	37.22 (10.28)
			MSE		
Main effects	74.05	74.83	69.75	64.33	59.54
+ 2-way interaction	85.14 (11.09)	80.80 (5.97)	82.55 (12.80)	76.38 (2.01)	70.86 (11.32)
+ 3-way interaction	85.73 (0.59)	80.96 (0.16)	83.68 (1.13)	78.39 (12.05)	75.87 (5.01)
			95% CI Coverage		
Main effects	12.13	24.62	57.20	42.97	66.48
+ 2-way interaction	42.32 (30.19)	77.72 (53.10)	90.52 (33.32)	71.89 (28.92)	87.50 (21.02)
+ 3-way interaction	67.89 (25.57)	80.86 (3.14)	97.10 (6.58)	88.28 (16.39)	91.37 (3.87)

The covariate indicating if at least one covariate of X_1, X_2, or X_3 showed to have a percent explained variance greater than 1% and to be statistically significant ($p \leq 0.0001$) in all five model dependent variables for the intercept bias, MSE, and 95% CI coverage (Table 4.2). Bias in the intercept was associated ($p \leq 0.0001$) with sample size and variance in the models for univariate sensitivity, univariate specificity, Youden's (1950) index, and also with sensitivity in the bivariate models. For intercept bias, sensitivity mean was only significantly associated ($p \leq 0.0001$) in the Youden's (1950) index. Correlation was significantly associated ($p \leq 0.0001$) with bias in the intercept in Youden's (1950) index and sensitivity in the bivariate model. The two-way interaction between sensitivity mean and variance was significantly associated ($p \leq 0.0001$) with bias in the intercept in univariate sensitivity, Youden's (1950) index and sensitivity in the bivariate model.

For the ANOVA model of the intercept MSE, sample size and variance was statistically significant ($p \leq 0.0001$) in all five models. The mean of sensitivity was significantly associated ($p \leq 0.0001$) with MSE in the intercept in univariate sensitivity and sensitivity in the bivariate model.

When ANOVA was applied to the simulated data for 95% CI coverage of the intercept, there were a number of statistically significant covariates. The 95% CI coverage of the univariate sensitivity model was associated ($p \leq 0.0001$) with the two-way interaction of the mean of sensitivity and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3 and the two-way interaction of variance and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3. The 95% CI coverage of the univariate specificity model was associated ($p \leq 0.0001$) with the variance and the two-way interaction of variance and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3. The 95% CI coverage of the Youden's (1950)

index model was associated ($p \leq 0.0001$) with sensitivity mean, variance, correlation, two-way interaction between sensitivity mean and variance, two-way interaction between variance and correlation, the two-way interaction of variance and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3, and the two-way interaction of correlation and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3. The 95% CI coverage of sensitivity in the bivariate model was associated ($p \leq 0.0001$) with the mean of sensitivity, correlation, the two-way interaction of sensitivity mean and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3, and the two-way interaction of correlation and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3. The 95% CI coverage of sensitivity in the bivariate model was associated ($p \leq 0.0001$) with the two-way interaction between sensitivity mean and correlation. The 95% CI coverage of specificity in the bivariate model was associated ($p \leq 0.0001$) with sample size, variance, two-way interaction between sensitivity mean and variance, and the two-way interaction of variance and the covariate indicating the use of at least one covariate of X_1, X_2, or X_3.

Table 4.2: Percent explained variance of the intercept bias, mean squared error (MSE), and 95% confidence interval (CI) coverage in the ANOVA model

Covariate	Univariate sensitivity	Univariate specificity	Univariate Youden's Index	Bivariate sensitivity	Bivariate specificity
			Bias		
N	**7.62**	**16.52**	**3.01**	**6.09**	1.58
Sensitivity mean	2.22	0.00	**1.34**	2.16	3.12
Variance	**15.40**	**19.44**	**31.93**	**18.45**	3.21
Correlation	0.00	0.01	**11.28**	**4.93**	2.07
Covariate	**31.76**	**31.64**	**12.62**	**22.81**	2.20
N*sensitivity mean	0.15	0.00	0.00	0.12	1.64
N*variance	0.67	0.69	0.38	0.77	0.14
N*correlation	0.00	0.00	0.05	0.03	1.02
N*covariate	0.98	1.32	0.50	1.13	0.84
Sensitivity mean*variance	**2.57**	0.00	**1.91**	**5.49**	1.22
Sensitivity mean*correlation	0.01	0.01	1.10	0.79	1.69
Sensitivity mean*covariate	2.12	0.00	0.14	1.68	0.46
Variance*correlation	0.00	0.00	**3.90**	0.41	3.14
Variance*covariate	0.79	0.51	0.40	0.46	0.47
Correlation*covariate	0.01	0.00	0.09	0.42	0.39
			MSE		
N	**5.74**	**12.04**	**4.55**	**4.26**	**9.20**
Sensitivity mean	**13.45**	0.00	1.98	**9.26**	1.03
Variance	**8.81**	**12.56**	**12.90**	**7.08**	**5.65**
Correlation	0.00	0.00	2.90	1.06	2.02
Covariate	**18.06**	**28.30**	**19.47**	**17.33**	**21.36**
N*sensitivity mean	**2.61**	0.00	0.49	1.47	0.51
N*variance	1.78	2.15	**2.63**	0.86	1.15
N*correlation	0.00	0.00	0.54	0.46	0.31
N*covariate	1.84	**2.84**	1.89	1.58	1.54
Sensitivity mean*variance	0.90	0.00	0.19	0.86	0.05
Sensitivity mean*correlation	0.00	0.00	0.08	1.22	0.21
Sensitivity mean*covariate	2.03	0.00	0.50	2.42	0.06
Variance*correlation	0.00	0.00	0.83	0.56	2.92
Variance*covariate	1.92	0.97	**4.61**	2.23	3.11
Correlation*covariate	0.00	0.00	1.04	0.39	1.44
			95% CI Coverage		
N	2.35	0.74	1.21	2.09	**4.83**
Sensitivity mean	0.84	0.20	**2.21**	**15.97**	**21.25**

Variance	0.48	**5.55**	**38.11**	0.01	**10.85**
Correlation	0.39	0.94	**19.56**	**25.24**	**10.22**
Covariate	0.01	**19.33**	**12.27**	**15.80**	**29.94**
N*sensitivity mean	0.89	0.36	0.16	0.22	1.04
N*variance	0.99	0.35	1.17	0.22	0.14
N*correlation	0.35	0.19	0.68	1.31	0.70
N*covariate	0.15	0.78	0.59	0.72	0.83
Sensitivity mean*variance	0.94	0.08	**1.98**	2.19	**7.94**
Sensitivity mean*correlation	0.41	0.43	0.64	**6.94**	21.25
Sensitivity mean*covariate	**13.68**	0.38	0.10	**5.01**	**4.11**
Variance*correlation	0.50	0.64	**12.15**	0.66	0.10
Variance*covariate	**9.95**	**49.85**	**10.91**	0.07	**2.98**
Correlation*covariate	2.35	0.03	**4.93**	**11.59**	**3.04**

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$

Tables 4.3 to 4.5 focus on the covariates that were associated with bias, MSE, and 95% CI coverage for the covariates X_1, X_2, and X_3. Overall, sample size and variance were shown to be statistically significant ($p \leq 0.0001$) and have a percent explained variance greater than 1% in all ANOVA models for X_1, X_2, and X_3 except bias in X_3.

Table 4.3: Percent explained variance of the X_1 coefficient bias, mean squared error (MSE), and 95% confidence interval (CI) coverage

Covariate	Univariate sensitivity	Univariate specificity	Univariate Youden's Index	Bivariate sensitivity	Bivariate specificity
			Bias		
N	9.31	**80.15**	**16.09**	**23.23**	**36.72**
Sensitivity mean	0.76	0.36	1.21	**21.09**	**8.63**
Variance	0.01	0.72	**40.53**	**13.05**	**19.48**
Correlation	2.52	1.25	**12.72**	**10.30**	**13.34**
N*sensitivity mean	5.91	1.35	0.24	**8.10**	0.52
N*variance	3.65	1.42	**10.36**	3.94	**4.07**
N*correlation	2.48	0.55	**2.55**	0.56	**1.83**
Sensitivity mean*variance	4.88	1.16	1.84	0.77	**3.50**
Sensitivity mean*correlation	5.44	0.05	1.43	3.99	0.17
Variance*correlation	3.06	0.49	**7.08**	**5.51**	**7.22**
			MSE		
N	**26.01**	**72.06**	**20.70**	**19.72**	**36.70**
Sensitivity mean	**35.86**	0.00	**6.50**	**33.02**	**2.80**
Variance	**30.38**	**24.57**	**51.96**	**25.30**	**28.51**
Correlation	0.04	0.00	**11.70**	**5.28**	**12.70**
N*sensitivity mean	**3.05**	0.00	**0.77**	2.25	0.63
N*variance	**2.51**	**3.00**	**4.92**	1.63	**3.71**
N*correlation	0.03	0.00	**1.05**	1.09	1.10
Sensitivity mean*variance	0.56	0.00	0.07	0.46	0.94
Sensitivity mean*correlation	0.08	0.00	0.25	2.12	0.66
Variance*correlation	0.04	0.00	**1.17**	1.15	**5.78**
			95% CI Coverage		
N	**17.20**	**63.48**	**15.96**	**22.03**	**56.85**
Sensitivity mean	**39.32**	0.04	**13.19**	**30.44**	0.50
Variance	**36.72**	**34.96**	**52.71**	**34.25**	**24.47**
Correlation	0.02	0.03	**12.69**	**2.07**	**11.86**
N*sensitivity mean	0.43	0.06	0.03	**0.84**	0.55
N*variance	0.38	**0.86**	**0.66**	**1.02**	**1.14**
N*correlation	0.04	0.01	0.02	0.08	**1.72**
Sensitivity mean*variance	**4.18**	0.02	**0.96**	**8.13**	0.18
Sensitivity mean*correlation	0.11	0.04	0.44	0.03	0.09
Variance*correlation	0.08	0.00	**1.84**	0.02	0.45

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$

Table 4.4: Percent explained variance of the X_2 coefficient bias, mean squared error (MSE), and 95% confidence interval (CI) coverage

Covariate	Univariate sensitivity	Univariate specificity	Univariate Youden's Index	Bivariate sensitivity	Bivariate specificity
			Bias		
N	**27.97**	8.01	**16.13**	**18.94**	**46.52**
Sensitivity mean	**14.37**	2.60	0.41	**21.82**	**5.32**
Variance	**18.82**	0.71	**39.20**	**16.25**	**9.29**
Correlation	0.13	1.77	**12.18**	**7.52**	**14.52**
N*sensitivity mean	**5.91**	1.65	0.03	**8.27**	1.19
N*variance	**10.16**	1.30	**9.93**	**6.38**	**3.93**
N*correlation	0.20	1.74	**3.21**	1.13	**2.02**
Sensitivity mean*variance	**5.79**	1.81	**2.20**	2.08	**2.76**
Sensitivity mean*correlation	0.41	3.20	1.50	2.70	0.61
Variance*correlation	0.24	0.79	**6.44**	**4.01**	**6.08**
			MSE		
N	**27.91**	**72.58**	**20.64**	**21.08**	**35.88**
Sensitivity mean	**34.30**	0.01	**6.43**	**30.35**	**3.81**
Variance	**29.55**	**23.05**	**51.81**	**25.00**	**26.70**
Correlation	0.00	0.01	**11.69**	**5.70**	**14.51**
N*sensitivity mean	**3.76**	0.01	**0.76**	**2.86**	0.74
N*variance	**3.07**	**2.58**	**4.89**	**1.68**	**3.27**
N*correlation	0.00	0.01	**1.07**	0.83	1.10
Sensitivity mean*variance	**0.78**	0.01	0.05	0.40	**1.67**
Sensitivity mean*correlation	0.00	0.03	0.26	**2.51**	0.69
Variance*correlation	0.00	0.02	**1.15**	1.17	**6.58**
			95% CI Coverage		
N	**20.18**	**60.40**	**16.27**	**23.79**	**50.28**
Sensitivity mean	**36.99**	0.16	**12.89**	**25.77**	**1.10**
Variance	**36.30**	**34.51**	**52.74**	**35.11**	**29.07**
Correlation	0.00	0.10	**12.82**	**2.25**	**11.89**
N*sensitivity mean	**0.30**	0.06	0.02	0.49	**0.72**
N*variance	**0.30**	**1.02**	**0.64**	**1.47**	**1.00**
N*correlation	0.00	0.08	0.01	0.30	**2.00**
Sensitivity mean*variance	**5.43**	0.09	**0.80**	**5.87**	0.07
Sensitivity mean*correlation	0.00	0.14	**0.49**	0.26	0.04
Variance*correlation	0.00	0.02	**1.76**	0.08	**1.18**

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$

Table 4.5: Percent explained variance of the X_3 coefficient bias, mean squared error (MSE), and 95% confidence interval (CI) coverage

Covariate	Univariate sensitivity	Univariate specificity	Univariate Youden's Index	Bivariate sensitivity	Bivariate specificity
			Bias		
N	0.02	1.55	6.45	0.41	**18.75**
Sensitivity mean	**31.06**	2.87	15.56	5.36	5.47
Variance	4.16	0.84	1.18	6.68	8.36
Correlation	1.96	2.57	0.52	10.10	6.52
N*sensitivity mean	0.16	2.72	1.98	1.44	5.58
N*variance	0.94	1.18	4.78	0.68	4.70
N*correlation	1.47	2.96	1.23	4.54	4.49
Sensitivity mean*variance	3.85	2.81	11.44	8.19	8.89
Sensitivity mean*correlation	1.01	6.36	8.42	8.61	5.65
Variance*correlation	2.26	2.70	1.68	1.04	1.97
			MSE		
N	**28.82**	**71.50**	**21.22**	**19.45**	**35.56**
Sensitivity mean	**33.76**	0.01	**6.23**	**30.22**	**3.24**
Variance	**29.43**	**24.61**	**51.73**	**25.07**	**27.71**
Correlation	0.00	0.01	**11.53**	**5.28**	**13.74**
N*sensitivity mean	**3.88**	0.01	**0.77**	2.21	0.79
N*variance	**3.15**	**3.12**	**4.97**	1.17	**2.93**
N*correlation	0.00	0.01	**1.04**	1.13	1.36
Sensitivity mean*variance	**0.74**	0.00	0.05	0.49	1.34
Sensitivity mean*correlation	0.00	0.02	0.30	2.23	1.01
Variance*correlation	0.00	0.00	**1.12**	1.09	**6.27**
			95% CI Coverage		
N	**33.05**	**69.99**	**18.45**	**28.26**	**54.34**
Sensitivity mean	**27.21**	0.68	**4.85**	**28.66**	**3.42**
Variance	**22.08**	**8.58**	**50.89**	**22.91**	**23.76**
Correlation	0.00	0.63	**9.59**	**4.73**	**11.80**
N*sensitivity mean	**7.93**	0.59	**2.34**	**4.65**	0.11
N*variance	**6.87**	0.58	**7.34**	**4.32**	0.04
N*correlation	0.00	0.64	0.68	0.03	0.21
Sensitivity mean*variance	**1.72**	0.61	0.59	**1.54**	0.63
Sensitivity mean*correlation	0.00	1.24	0.12	**1.78**	0.02
Variance*correlation	0.00	0.63	**3.27**	**1.06**	**3.73**

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$

The covariates significantly associated ($p \leq 0.0001$) with the bivariate model convergence percent were sample size, sensitivity mean, variance, use of at least one covariate of X_1, X_2, and X_3, two-way interaction of sample size and variance, two-way interaction of sensitivity mean and variance, two-way interaction of sensitivity mean and the use of at least one covariate of X_1, X_2, and X_3, and the interaction between the variance and the use of at least one covariate of X_1, X_2, and X_3 (Table 4.6). The variance of sensitivity and specificity contributed the most to the percent explained variance in the bivariate model convergence.

Table 4.6: Percent explained variance of the bivariate model convergence

Covariate	**Percent explained variance**
N	**2.75**
Sensitivity mean	**1.81**
Variance	**28.63**
Correlation	0.26
Covariate	**11.83**
N*sensitivity mean	0.25
N*variance	**3.23**
N*correlation	0.06
N*covariate	0.21
Sensitivity mean*variance	**8.49**
Sensitivity mean*correlation	1.09
Sensitivity mean*covariate	**3.64**
Variance*correlation	1.31
Variance*covariate	**5.63**
Correlation*covariate	0.04

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$

4.2 Numeric Study

Results for the numeric study are described in two sections, the first section focuses on the descriptive analyses and the second focuses on the results for the statistical models. The first section describes the frequencies and means of the diagnostic validity measures for each of the case definition characteristics (Tables 4.7 and 4.8). The statistical modeling section tests these relationships (Tables 4.9 to 4.13).

4.2.1 Descriptive Analyses

This section describes the means and standard deviations of sensitivity, specificity, and Youden's index stratified by each case definition characteristic. The frequencies of case definition characteristics and correlations among these characteristics are also described.

At least one hospital discharge record (0 versus 1+) was included as a characteristic in two-thirds of the case definitions (64.9%) and showed to lower the mean estimate of sensitivity compared to when at least one hospital discharge record was not used (Table 4.7). At least one diagnosis in ER records (0 versus 1+) was only included as a characteristic in 10 case definitions and showed to decrease the means of the diagnostic validity measures. At least one diagnosis in physician claims (0 versus 1+) was used included as a characteristic in almost all case definitions (94.6%), and resulted in an increase in the average sensitivity and Youden's (1950) index but a decrease in the average specificity compared to when at least one diagnosis in physician claims was not used.

The number of physician claims had four categories: none, 1+, 2+, and 3+ (at least 3 physician claims to an unlimited maximum number). Increasing the number of physician claims from 1+ to 3+ showed a large increase in the mean of specificity. The length of the observation period to identify the physician diagnoses had five categories: never, 1 year, 2 years, 3-5 years, and ever (no time restriction). Increasing the observation time decreased specificity and Youden's (1950) index from the number of incorrectly identified positive cases increasing.

At least one specialist diagnosis was included as a characteristic in half of the case definitions tested (51.4%), and resulted in a small increase in the average values for all diagnostic validity measures. When increasing the number of specialist diagnoses from 1+ to 3+, the means of the diagnostic validity measures showed little change. A time separation between

two physician diagnoses was included as a characteristic in 14 case definitions. It showed to decrease the means of the diagnostic validity measures with a large decrease in the mean of sensitivity.

Exclusion criteria A resulted in the exclusion of individuals with 2+ physician diagnosis codes, using ICD-9, with a different rheumatology diagnosis to an RA diagnosis including osteoarthritis (715), gout (274, 712), polymyalgia rheumatic (725), other seronegative spondyloarthropathy (721), ankylosing spondylitis (720), psoriasis (696), synovitis/tenosynovitis/bursitis (727), connective tissue disorder (710), vasculitis (446), and others (716, 718, 727, 728, 729, 781). Exclusion criteria B excluded individuals whose RA diagnosis was not confirmed by a specialist. Exclusion criteria A and B (yes versus no) were not used frequently in the case definitions (6.8% and 5.4%, respectively). They resulted in a decrease in the average values of all diagnostic validity measures.

Of the 91 case definitions for the 65+ age group, 17 case definitions required 1+ prescription of a steroid, DMARD, or biological agent to ascertain cases and 17 case definitions required 1+ prescription of DMARD or biological agent. Using 1+ prescription of a steroid, DMARD, or biological agent showed no change on the means of the diagnostic validity measures. Using 1+ prescription of DMARD or biological agent to ascertain RA cases increased the mean of specificity.

Table 4.7: Characteristics of the rheumatoid arthritis (RA) case definitions (*n*=148)

Characteristic	Frequency (%)	Sensitivity Mean (SD)	Specificity Mean (SD)	Youden's Index Mean (SD)
Cohort				
20 + years	57 (38.51)	90.87 (17.06)	82.19 (7.48)	73.06 (15.28)
65+ years	91 (61.49)	91.30 (15.16)	86.10 (7.35)	77.40 (13.86)
# of hospital diagnoses				
1+	96 (64.86)	88.10 (18.96)	85.10 (8.01)	73.20 (17.02)
0	52 (35.14)	96.74 (2.42)	83.67 (6.82)	80.41 (5.87)
# of ER diagnoses				
1+	10 (6.76)	71.37 (34.05)	81.30 (14.92)	52.67 (22.08)
0	138 (93.24)	92.57 (12.77)	84.83 (6.84)	77.40 (12.36)
# of physician diagnoses				
1+	140 (94.59)	94.59 (6.56)	83.89 (7.22)	78.48 (9.02)
0	8 (5.41)	30.73 (4.38)	96.94 (1.07)	27.66 (5.01)
# of physician diagnoses				
0	8 (5.41)	30.73 (4.38)	96.94 (1.07)	27.66 (5.01)
1+	16 (10.13)	98.36 (2.42)	71.34 (9.71)	69.71 (8.69)
2+	64 (43.24)	93.52 (9.20)	83.15 (4.80)	76.67 (10.64)
3+	60 (40.54)	94.72 (2.10)	88.03 (3.66)	82.75 (2.95)
Diagnosis observation time				
Never	8 (5.41)	30.73 (4.38)	96.94 (1.07)	27.66 (5.01)
1 year	38 (25.68)	94.86 (2.84)	87.51 (5.42)	82.37 (3.94)
2 years	40 (27.03)	94.55 (7.27)	85.02 (4.38)	79.57 (8.49)
3-5 years	44 (29.73)	93.17 (8.66)	84.14 (4.48)	77.31 (10.60)
Ever	18 (12.16)	97.59 (2.88)	73.12 (10.57)	70.71 (8.83)
# of specialist diagnoses				
1+	76 (51.35)	96.22 (2.65)	87.27 (3.75)	83.50 (2.79)
0	72 (48.65)	85.77 (21.37)	81.77 (9.46)	67.54 (17.21)
# of specialist diagnoses				
0	72 (48.65)	85.77 (21.37)	81.77 (9.46)	67.54 (17.21)
1+	40 (27.03)	96.36 (2.88)	86.49 (4.46)	82.85 (2.93)
2+	24 (16.22)	96.93 (2.09)	87.59 (2.61)	84.52 (2.43)
3+	12 (8.11)	94.34 (2.11)	89.26 (2.12)	83.60 (2.49)
Time separation between physician diagnoses				
Yes	14 (9.46)	79.77 (11.08)	80.53 (2.91)	60.30 (10.78)
No	134 (90.54)	92.33 (15.84)	85.02 (7.84)	77.34 (13.94)
Exclusion criteria A				
Yes	10 (6.76)	73.18 (2.58)	80.77 (1.27)	53.95 (2.56)
No	138 (93.24)	92.44 (15.63)	84.87 (7.81)	77.31 (13.74)
Exclusion criteria B				
Yes	8 (5.41)	73.70 (2.61)	80.46 (1.22)	54.16 (2.79)
No	140 (94.59)	92.13 (15.72)	84.83 (7.76)	76.97 (13.95)

# of RX				
1+	17 (11.49)	91.12 (15.45)	84.54 (8.38)	75.75 (13.54)
0	74 (50.00)	91.32 (15.20)	86.46 (7.11)	77.79 (14.00)
N/A	57 (38.51)	90.88 (17.06)	82.19 (7.48)	73.06 (15.28)
# of DMARD or biological prescriptions				
1+	17 (11.49)	88.66 (14.83)	91.24 (3.70)	79.90 (13.48)
0	74 (50.00)	91.91 (15.27)	84.92 (7.49)	76.83 (13.97)
N/A	57 (38.51)	90.88 (17.06)	82.19 (7.48)	73.06 (15.28)

Note: RX indicates either a steroid, disease-modifying antirheumatic drug (DMARD), or biological agent was prescribed

Table 4.8 displays spearman correlations for the case definition characteristics. The following case definition characteristics were highly correlated: exclusion criteria A and B (0.89; $p<0.0001$), exclusion criteria A and using a time separation between physician claims (0.83; $p<0.0001$), exclusion criteria B and using a time separation between physician claims (0.74; $p<0.0001$), and the binary number of specialist diagnoses and the ordinal number of specialist diagnoses (0.84; $p<0.0001$). These combinations of case definition characteristics were not included in the same model due to high collinearity.

Table 4.8: Correlations (standard errors) among case definition characteristics

	# hospital diagnoses	# ER diagnoses	physician claims	# physician claims	Observation time	specialist	# specialist diagnoses	Separation	Exclusion A	Exclusion B	RX	DMARD or biological agent
# hospital diagnoses	1.00											
# ER diagnoses	**0.20** **(0.03)**	1.00										
Physician claims	**-0.18** **(0.03)**	**-0.41** **(0.15)**	1.00									
# physician claims	-0.11 (0.08)	**-0.43** **(0.06)**	**0.42** **(0.07)**	1.00								
Observation time	-0.11 (0.08)	-0.07 (0.12)	**0.40** **(0.06)**	-0.13 (0.10)	1.00							
Specialist	-0.15 (0.08)	**-0.28** **(0.04)**	**0.25** **(0.04)**	**0.37** **(0.07)**	-0.10 (0.08)	1.00						
# specialist diagnoses	**-0.18** **(0.08)**	**-0.26** **(0.04)**	**0.23** **(0.04)**	**0.42** **(0.07)**	-0.11 (0.08)	**0.93** **(0.01)**	1.00					
Separation	0.14 (0.06)	-0.09 (0.02)	0.08 (0.02)	-0.15 (0.04)	**0.25** **(0.06)**	**-0.33** **(0.05)**	**-0.31** **(0.04)**	1.00				
Exclusion A	**0.20** **(0.03)**	-0.07 (0.02)	0.06 (0.02)	-0.12 (0.04)	**0.19** **(0.06)**	**-0.28** **(0.04)**	**-0.26** **(0.04)**	**0.83** **(0.08)**	1.00			
Exclusion B	**0.18** **(0.03)**	-0.06 (0.02)	0.06 (0.01)	-0.11 (0.03)	**0.19** **(0.06)**	**-0.25** **(0.04)**	**-0.23** **(0.04)**	**0.74** **(0.09)**	**0.89** **(0.07)**	1.00		
RX	0.15 (0.09)	-0.01 (0.10)	-0.01 (0.11)	-0.04 (0.10)	0.05 (0.11)	-0.13 (0.10)	-0.18 (0.08)	-0.05 (0.09)	-0.12 (0.03)	-0.10 (0.03)	1.00	
DMARD or biological agent	0.15 (0.09)	-0.01 (0.10)	-0.01 (0.11)	-0.04 (0.10)	0.05 (0.11)	-0.13 (0.10)	-0.18 (0.08)	-0.05 (0.09)	-0.12 (0.03)	-0.10 (0.03)	**-0.23** **(0.04)**	1.00

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.05$; RX indicates either a steroid, disease-modifying antirheumatic drugs (DMARD), or biological agent was prescribed

4.2.2 Statistical Model

This section describes the unadjusted and adjusted (i.e., multivariable) models that were applied to the data to test the relationships between the diagnostic validity measures and the case definition characteristics. Before fitting the unadjusted models, null models were fit to the data (Table 4.9). Fit statistics were produced for the univariate models of sensitivity, specificity, and Youden's (1950) index when both age groups were used (i.e., 20+ population and 65+ population) and when only the case definitions applied to the 65+ population were used. These two groups were looked at because when the case definition characteristic of the number of prescriptions for DMARD or biological agent was used it limits the population to those 65+ years of age. Table 4.10 fits null models to the bivariate model with case definitions applied to both age groups (n=148) and with only the case definitions applied to the 65+ age group (n=91). The null models were used to assess the fit of the multivariable models by comparing the AIC, BIC, and -2 log likelihood between the null models with the corresponding multivariable models.

Table 4.9: Logit estimates and standard errors (SEs) for null univariate models of sensitivity, specificity, and Youden's index with all case definitions and the case definitions where the prescription data was available

	All age groups (n=148)			65+ group (n=91)		
	Sensitivity	**Specificity**	**Youden's index**	**Sensitivity**	**Specificity**	**Youden's index**
Null model	**2.20 (0.12)**	**1.70 (0.05)**	**1.09 (0.06)**	**2.25 (0.16)**	**1.82 (0.06)**	**1.19 (0.07)**
Fit Statistics						
AIC	-483.43	-357.84	-198.69	-319.94	-229.15	-133.68
BIC	-477.44	-351.84	-192.70	-314.92	-224.13	-128.65
-2 Log-Likelihood	-487.43	-361.84	-202.69	-323.94	-233.15	-137.68

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$; AIC - Akaike information criterion; BIC - Bayesian information criterion

Table 4.10: Logit estimates and standard errors (SEs) for the null bivariate model of sensitivity and specificity for all age groups and the 65+ age group

	All age groups (n=148)		65+ group (n=91)	
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**
Null model	**3.11 (0.14)**	**1.76 (0.05)**	**3.14 (0.18)**	**1.88 (0.06)**
Fit Statistics				
AIC	-925.96		-596.64	
BIC	-910.97		-584.09	
-2 Log-Likelihood	-935.96		-606.64	

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$; AIC - Akaike information criterion; BIC - Bayesian information criterion

The variables were statistically significant in the univariate sensitivity model were: number of physician diagnoses (p<0.0001), length of observation period to identify the physician diagnoses (p<0.0001), time separation between two physician diagnoses (p<0.0001), exclusion criteria A and B (p<0.0001), and the number of prescriptions for a DMARD or biological agent (p=0.0030) (Table 4.11). The case definition characteristics associated with specificity in the univariate model were: number of physician diagnoses (p<0.0001), length of observation period to identify the physician diagnoses (p<0.0001), number of specialist diagnoses (p=0.0062), time separation between two physician diagnoses (p=0.0056), and the number of prescriptions for DMARD or biological agent (p=0.0027). The case definition characteristics that were associated with Youden's (1950) index were: number of diagnoses in ER records (p<0.0001), number of physician diagnoses (p<0.0001), length of observation period to identify the physician diagnoses (p<0.0001), number of specialist visits (p<0.0001), a time separation between two physician diagnoses (p<0.0001), and exclusion criteria A and B (p<0.0001).

The bivariate unadjusted models of sensitivity and specificity produced three statistically significant case definition characteristics: number of physician diagnoses (p<0.0001), length of observation period to identify the physician diagnoses, and number of specialist diagnoses (sensitivity p=0.0005; specificity p<0.0001). In addition, sensitivity in the bivariate unadjusted models were associated with: number of diagnoses in hospital discharge records (p=0.0007), a time separation between two physician diagnoses (p<0.0001), and exclusion criteria A and B (p<0.0001). Specificity in the bivariate unadjusted models were associated with the number of prescriptions for DMARD or biological agent (p=0.0011).

Table 4.11: Logit estimates and standard errors (SEs) for unadjusted univariate models of sensitivity, specificity, Youden's index, and sensitivity and specificity in the bivariate model (*n*=148)

Characteristics	Univariate sensitivity	Univariate specificity	Youden's index	Bivariate sensitivity	Bivariate specificity
# of hospital diagnoses					
Intercept	**2.46 (0.17)**	**1.59 (0.80)**	**1.28 (0.10)**	**3.72 (0.23)**	**1.65 (0.08)**
1+	-0.38 (0.17)	0.18 (0.09)	-0.28 (0.11)	**-0.95 (0.27)**	0.18 (0.10)
# of ER diagnoses					
Intercept	**2.23 (0.12)**	**1.70 (0.05)**	**1.18 (0.05)**	**3.18 (0.14)**	**1.78 (0.05)**
1+	-0.33 (0.32)	0.05 (0.18)	**-1.08 (0.18)**	-1.06 (0.51)	-0.14 (0.18)
# of physician diagnoses					
Intercept	**-0.74 (0.21)**	**3.09 (0.27)**	**-0.93 (0.15)**	-0.78 (0.37)	**3.19 (0.26)**
1+	**3.65 (0.24)**	**-1.44 (0.27)**	**2.22 (0.16)**	**4.03 (0.38)**	**-1.51 (0.27)**
# of physician diagnoses					
Intercept	**-0.76 (0.19)**	**3.25 (0.23)**	**-0.94 (0.13)**	-0.83 (0.35)	**3.27 (0.22)**
0 (reference)	0	0	0	0	0
1+	**4.72 (0.33)**	**-2.31 (0.24)**	**1.78 (0.16)**	**5.50 (0.48)**	**-2.33 (0.24)**
2+	**3.95 (0.23)**	**-1.66 (0.23)**	**2.15 (0.14)**	**4.17 (0.38)**	**-1.66 (0.23)**
3+	**3.26 (0.22)**	**-1.27 (0.23)**	**2.46 (0.15)**	**3.80 (0.38)**	**-1.29 (0.23)**
Diagnosis observation time					
Intercept	**2.54 (0.14)**	**1.96 (0.07)**	**1.50 (0.07)**	**3.05 (0.19)**	**1.96 (0.07)**
Never	**-3.29 (0.25)**	**1.24 (0.25)**	**-2.44 (0.16)**	**-3.89 (0.41)**	**1.28 (0.24)**
1 year (reference)	0	0	0	0	0
2 years	0.47 (0.20)	-0.25 (0.10)	-0.15 (0.10)	0.26 (0.26)	-0.23 (0.10)
3-5 years	0.37 (0.19)	**-0.31 (0.09)**	**-0.25 (0.10)**	0.06 (0.26)	**-0.29 (0.09)**
Ever	**1.19 (0.27)**	**-0.92 (0.11)**	**-0.61 (0.12)**	**1.27 (0.36)**	**-0.93 (0.11)**
# of specialist diagnoses					
Intercept	**2.06 (0.14)**	**1.58 (0.06)**	**0.74 (0.06)**	**2.71 (0.18)**	**1.61 (0.07)**
1+	0.28 (0.17)	**0.25 (0.09)**	**0.74 (0.10)**	**0.86 (0.26)**	**0.32 (0.09)**
# of specialist diagnoses					

Intercept	**2.07 (0.14)**	**1.58 (0.06)**	**0.74 (0.06)**	**2.71 (0.18)**	**1.61 (0.06)**
1+	0.34 (0.20)	0.20 (0.11)	**0.71 (0.12)**	**0.96 (0.31)**	0.26 (0.11)
2+	0.38 (0.24)	0.26 (0.13)	**0.81 (0.14)**	**1.06 (0.37)**	0.35 (0.13)
3+	-0.18 (0.31)	0.39 (0.18)	**0.75 (0.19)**	0.20 (0.48)	**0.48 (0.18)**
Time separation between physician diagnoses					
Intercept	**2.35 (0.12)**	**1.74 (0.05)**	**1.18 (0.06)**	**3.39 (0.14)**	**1.81 (0.05)**
Yes	**-1.09 (0.26)**	**-0.40 (0.14)**	**-0.77 (0.16)**	**-1.81 (0.43)**	-0.38 (0.16)
Exclusion criteria A					
Intercept	**2.39 (0.12)**	**1.73 (0.05)**	**1.18 (0.05)**	**3.39 (0.14)**	**1.82 (0.05)**
Yes	**-1.59 (0.29)**	-0.38 (0.17)	**-1.13 (0.18)**	**-2.37 (0.49)**	-0.40 (0.20)
Exclusion criteria B					
Intercept	**2.35 (0.12)**	**1.73 (0.05)**	**1.16 (0.06)**	**3.32 (0.14)**	**1.81 (0.05)**
Yes	**-1.53 (0.32)**	-0.39 (0.18)	**-1.01 (0.21)**	**-2.37 (0.55)**	-0.39 (0.21)
# of RX					
Intercept	**2.37 (0.17)**	**1.84 (0.07)**	**1.21 (0.08)**	**3.21 (0.20)**	**1.92 (0.07)**
1+	-0.55 (0.27)	-0.11 (0.15)	-0.13 (0.17)	-0.38 (0.44)	-0.18 (0.16)
# of DMARD or biological agent prescriptions					
Intercept	**2.44 (0.17)**	**1.74 (0.06)**	**1.16 (0.08)**	**3.33 (0.19)**	**1.78 (0.07)**
1+	**-0.80 (0.26)**	**0.50 (0.16)**	0.15 (0.18)	-0.98 (0.42)	**0.54 (0.16)**

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.01$; RX indicates either a steroid, disease-modifying antirheumatic drugs (DMARD), or biological agent was prescribed; AIC - Akaike information criterion; BIC - Bayesian information criterion

The adjusted univariate models of sensitivity, specificity, and Youden's (1950) index were fit including all variables that were statistically significant in the unadjusted models. This resulted in two models for each dependent variable based on the inclusion and exclusion of the case definition characteristic indicating the number of prescriptions of DMARD or biological agent prescriptions. The first includes all case definitions and doesn't include the prescription characteristics. The second includes the case definitions only applied to the 65+ population and includes the prescription characteristics. The fit statistics show that all models presented in Table 4.12 provide better fit than the null models.

When all case definitions were considered, the univariate model of sensitivity showed that using 1+ physician diagnosis resulted in a statistically significant increase in sensitivity ($p<0.0001$), increasing the length of observation time to unlimited compared to 1 year was associated with a statistically significant increase in sensitivity ($p<0.0001$), and requiring a separation period between physician diagnoses significantly decreased sensitivity ($p<0.0001$) (Table 4.12). Similar patterns were found for the models for the case definitions from the 65+ years, except that increasing the length of observation time from 1 year was no longer statistically significant. This may be from 65+ year olds being more likely to attend physician visits in a timely manner or they may have more RA symptoms which require more visits compared to younger individuals diagnosed with RA.

When case definitions applied to all age groups (20+ and 65+) were considered, the univariate model of specificity showed that using 1+ physician diagnosis was associated with a statistically significant decrease in specificity ($p<0.0001$), increasing the observation time from 1 year to 3-5 years significantly decreased specificity ($p<0.0001$), and requiring 1+ specialist diagnosis significantly increased specificity ($p<0.0001$) (Table 4.12). When only considering the

case definitions for the 65+ age group, the results showed similar patterns. A DMARD or biological agent prescription was associated with a statistically significant increase in specificity (p<0.0001).

The initial univariate model for Youden's (1950) index included the case definition characteristics of 1+ diagnosis in ER records and physician diagnosis observation time. Neither characteristic was statistically significant and model fit improved when they were removed from the model (data not shown). When all case definitions were considered, the univariate model of Youden's (1950) index showed that using 1+ physician diagnosis to ascertain cases increased Youden's (1950) index (p<0.0001), requiring 1+ specialist diagnosis significantly increased Youden's (1950) index (p<0.0001), and requiring a separation period between diagnoses significantly decreased Youden's (1950) index (p<0.0001) (Table 4.12). When only considering the case definitions for the 65+ population, they showed similar patterns. Using a DMARD or biological agent prescription to ascertain RA cases resulted in a significant increase in Youden's (1950) index (p<0.0001).

Table 4.12: Logit estimates and standard errors (SEs) for adjusted univariate models of sensitivity, specificity, and Youden's index with all case definitions and the case definitions where the prescription data was available

	All age groups (n=148)			**65+ age group (n=91)**		
Characteristics	**Univariate sensitivity**	**Univariate specificity**	**Youden's index**	**Univariate sensitivity**	**Univariate specificity**	**Youden's index**
Intercept	**-0.79 (0.13)**	**3.32 (0.20)**	**-0.96 (0.09)**	-0.55 (0.18)	**3.57 (0.17)**	**-0.86 (0.09)**
# of physician diagnoses						
0 (reference)	0	0	0	0	0	0
1+	**4.06 (0.37)**	**-2.27 (0.24)**	**1.72 (0.10)**	**3.53 (0.47)**	**-2.27 (0.19)**	**1.65 (0.11)**
2+	**4.26 (0.22)**	**-1.65 (0.21)**	**2.21 (0.10)**	**4.14 (0.30)**	**-1.90 (0.17)**	**2.11 (0.10)**
3+	**3.14 (0.18)**	**-1.36 (0.21)**	**2.29 (0.10)**	**3.01 (0.25)**	**-1.60 (0.17)**	**2.17 (0.10)**
Diagnosis observation time						
Never	0	0		0	0	
1 year (reference)	0	0		0	0	
2 years	0.58 (0.16)	-0.27 (0.08)		0.49 (0.22)	**-0.32 (0.05)**	
3-5 years	0.55 (0.16)	**-0.31 (0.08)**		0.57 (0.22)	**-0.37 (0.05)**	
Ever	**1.38 (0.33)**	-0.21 (0.12)		1.47 (0.40)	**-0.54 (0.08)**	
# of specialist diagnoses		**0.34 (0.06)**			**0.47 (0.04)**	
# of specialist diagnoses						
0 (reference)			0			0
1+			**0.34 (0.06)**			**0.38 (0.06)**
2+			**0.40 (0.07)**			**0.58 (0.09)**
3+			0.29 (0.10)			**0.47 (0.12)**
Time separation between physician diagnoses	**-2.58 (0.19)**	0.02 (0.10)	**-0.80 (0.08)**	**-2.53 (0.29)**	0.11 (0.07)	**-0.69 (0.08)**
# of DMARD or biological agent prescriptions				-0.69 (0.21)	**0.77 (0.06)**	**0.29 (0.06)**

Fit Statistics						
AIC	-704.74	-524.32	-491.54	-440.36	-441.91	-333.67
BIC	-677.77	-494.34	-464.57	-415.25	-414.29	-308.56
-2 Log-Likelihood	**-722.74**	**-544.32**	**-509.54**	**-460.36**	**-463.91**	**-353.67**

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$; DMARD - disease-modifying antirheumatic drugs; AIC - Akaike information criterion; BIC - Bayesian information criterion

The bivariate model of sensitivity and specificity was fit using all statistically significant variables in the unadjusted models. Two models were fit to the data: The first included case definitions for all age groups and the second included case definitions for the 65+ age group. When the model was fit to the data, the number of hospital records with an RA diagnosis was not associated with these validity estimates (data not shown).

Table 4.13 shows the best-fitting bivariate models. When all case definitions were included, using 1+ physician diagnosis was associated with a statistically significant increase in sensitivity ($p<0.0001$) and a statistically significant decrease in specificity ($p<0.0001$). Including a time separation between physician diagnoses decreased sensitivity ($p<0.0001$) and has no significant change on specificity. When limited to the case definitions applied to the 65+ age group, the significant characteristics in all age groups remained significant. Additionally, increasing the number of observation years from 1 year significantly decreased specificity ($p<0.0001$), increasing the number of diagnoses by a specialist significantly increased specificity ($p<0.0001$), and using 1+ DMARD or biological agent prescription decreased sensitivity ($p<0.0001$) and increased specificity ($p<0.0001$). Based on the model fit statistics, both models resulted in better fit when compared to the null models.

Table 4.13: Logit estimates and standard errors for the adjusted bivariate models of sensitivity and specificity for all case definitions and the case definitions where the prescription data was available

	All age groups (n=148)		**65+ group (n=91)**	
Characteristics	**Bivariate sensitivity**	**Bivariate specificity**	**Bivariate sensitivity**	**Bivariate specificity**
Intercept	**-0.81 (0.17)**	**3.19 (0.25)**	-0.43 (0.41)	**3.55 (0.18)**
# of physician diagnoses				
0 (reference)	0	0	0	0
1+	**4.14 (0.42)**	**-2.16 (0.30)**	**4.33 (0.71)**	**-2.24 (0.21)**
2+	**4.23 (0.26)**	**-1.56 (0.26)**	**4.21 (0.50)**	**-1.88 (0.19)**
3+	**3.17 (0.24)**	**-1.27 (0.26)**	**2.83 (0.50)**	**-1.60 (0.19)**
Diagnosis observation time				
Never	0	0	0	0
1 year (reference)	0	0	0	0
2 years	0.59 (0.17)	-0.26 (0.10)	0.78 (0.30)	**-0.32 (0.06)**
3-5 years	0.58 (0.17)	-0.29 (0.10)	0.82 (0.29)	**-0.36 (0.06)**
Ever	1.45 (0.38)	-0.21 (0.17)	1.20 (0.52)	**-0.54 (0.09)**
# of specialist diagnoses				
0 (reference)	0	0	0	0
1	0.09 (0.18)	0.34 (0.09)	0.32 (0.27)	**0.42 (0.05)**
2	0.25 (0.21)	0.32 (0.11)	0.73 (0.37)	**0.56 (0.07)**
3	-0.00 (0.24)	0.31 (0.16)	0.18 (0.46)	**0.52 (0.10)**
Time separation between physician diagnoses	**-2.62 (0.23)**	0.03 (0.13)	**-2.66 (0.43)**	0.11 (0.08)
# of DMARD or biological agent prescriptions			**-1.33 (0.27)**	**0.77 (0.06)**
Fit Statistics				
AIC	-1202.56		-843.79	
BIC	-1127.63		-776.00	
-2 Log-Likelihood	**-1252.56**		**-897.79**	

Note: Values in bold face font indicate estimates that were statistically significant at $\alpha = 0.0001$; DMARD - disease-modifying antirheumatic drugs; AIC - Akaike information criterion; BIC - Bayesian information criterion

CHAPTER 5 – DISCUSSION AND CONCLUSIONS

5.1 Summary

The performance of three regression models, which were used to test for differences in the accuracy of case definitions form diagnostic validation studies, were compared: (a) univariate fixed-effects models independently applied to sensitivity and specificity estimates, (b) bivariate mixed-effects model applied to estimates of sensitivity and specificity, and (c) univariate fixed-effects model applied to a summary measure of sensitivity and specificity. A numeric example using a published RA validation study was used to demonstrate the methods.

Model performance was assessed using bias, MSE, and 95% CI coverage of the model intercept. When no covariates were present, the model intercepts represented the mean estimates of sensitivity, specificity, or Youden's (1950) index. The simulation results showed the univariate model intercepts of sensitivity and specificity had less bias, better accuracy as evaluated by MSE, and better coverage of the true population intercepts compared to the sensitivity intercept and specificity intercept in the bivariate model, for the conditions that were investigated. Additionally, the univariate models of sensitivity and specificity converged for all simulation conditions, while the bivariate model converged for 81% of the simulation conditions. The univariate model intercept of sensitivity had slightly less bias, better accuracy as evaluated by MSE, and better coverage of the true population intercepts compared to the univariate model intercept of Youden's (1950) index. The univariate model intercept for specificity had smaller bias and larger 95% CI coverage than the univariate model intercept of Youden's (1950) index; however, the univariate model intercept of Youden's (1950) index had smaller MSE.

Sample size and the magnitude of variability in sensitivity and specificity estimates were the simulation parameters significantly associated with bias and error of the intercepts in all

models. Sensitivity mean and correlation between sensitivity and specificity were the simulation parameters associated with 95% CI coverage of the intercepts of Youden's (1950) index, sensitivity in the bivariate model, and specificity in the bivariate model. Variance of sensitivity and specificity was a simulation parameter associated with 95% CI coverage of the intercepts of specificity in the univariate model, Youden's (1950) index and specificity in the bivariate model. The simulation parameters associated with convergence of the bivariate model were sample size, sensitivity mean, and variance of sensitivity and specificity.

In the numeric example, the three statistical models were applied in a secondary analysis of 148 case definitions, which were applied to administrative health data to ascertain cases of RA. Based on the results of the univariate model, increasing sensitivity was associated with increasing the number of physician diagnoses, increasing the number of observation years, and not requiring a time separation between physician diagnoses. Based on the results of the univariate model, increasing specificity was associated with decreasing the number of physician diagnoses, decreasing the number of observation years, requiring 1+ specialist diagnosis, and requiring 1+ prescription for a DMARD or biological agent. Based on the univariate model results, increasing Youden's (1950) index was associated with increasing the number of physician diagnoses, requiring 1+ specialist diagnosis, not requiring a time separation between physician diagnoses, and requiring 1+ prescription for a DMARD or biological agent. For the bivariate model, increasing sensitivity was associated with increasing the number of physician diagnoses, not requiring a time separation between physician diagnoses, and requiring at least one prescription for a DMARD or biological agent. For the bivariate model, increasing specificity was associated with decreasing the number of physician diagnoses, decreasing the

number of observation years, requiring 1+ specialist diagnosis, and requiring 1+ prescription for a DMARD or biological agent.

5.2 Discussion

There are no established guidelines to evaluate model performance in simulation studies, although recommendations on the acceptable levels of bias and 95% CI coverage have been described. Bias greater than $0.5SE(\hat{\beta})$ (Burton, Altman, Royston, & Holder, 2006; Schafer & Graham, 2002) and $2SE(\hat{\beta})$ (Burton et al., 2006; Sinharay, Stern, & Russel, 2001) have both been used as indicators of poor model performance. Burton et al. (2006) recommended that acceptable coverage be quantified as (Tang, Song, Belin, & Unützer, 2005):

$$p \mp Z_{1-(\alpha/2)}\sqrt{p(1-p)/B}, \tag{5.1}$$

where p is the $(1-\alpha)$ CI level, $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution, and B is the number of simulations. When 95% CI coverage falls above the upper bound the Type 2 error rate is too high (i.e., true differences are not detected) and when 95% CI coverage falls below the lower bound the Type 1 error rate is too high (i.e., false positive differences occur) (Burton et al., 2006; Tang et al., 2005).

Using these recommendations, the univariate model performed well if bias in the intercepts (i.e., means) was less than 8.0% and 95% CI coverage in the intercept was between 94.0% and 96.0%. Bias in the intercept of univariate models of sensitivity and specificity were below 8.0% for all scenarios. For the univariate models of sensitivity and specificity, 74.3% and 61.8% of the simulation scenarios fell within the 95% CI coverage bounds, respectively.

Bias in the intercept for the univariate model of Youden's (1950) index was below 8.0% in almost all simulation scenarios. Overall, only one-third of the simulation scenarios fell within

the 95% CI coverage bounds, indicating that the model was producing too many Type 1 or Type 2 errors.

Using these recommendations, performance of the sensitivity and specificity intercepts from the bivariate model was acceptable if bias was less than 25.0% and 95% CI coverage was between 94.0% and 96.0%. Bias in the bivariate sensitivity intercept was consistently less than this criterion. Bias in the bivariate specificity intercept was less than this criterion in 99% of the scenarios. Overall, 43.8% and 9.7% of the simulation scenarios fell in the acceptable range of CI coverage for sensitivity and specificity, respectively. The simulation parameters that decreased the CI coverage were a large sample size, large mean of sensitivity, large sensitivity and specificity variance, and large correlation between sensitivity and specificity. The combination of these parameters increased the number of observations reaching the maximum sensitivity value and maximum specificity value leading to an increased mean and smaller standard error. Therefore, the true population mean was more likely to fall outside of the estimated 95% CI and decrease the coverage.

The bivariate model failed to converge in one fifth of the simulation conditions. Burton et al. (2006) noted that if many simulations fail, the estimates may be biased or if the estimates are unbiased, the overall results may be influenced by selection bias (i.e., the simulation parameters or combinations of parameters may influence the scenarios that converge). The number of simulations was large enough to ensure the desired level of accuracy in bias estimates (Burton et al., 2006). Thus, the bivariate model estimates should not be biased from the non-converged simulations.

The univariate model of Youden's (1950) index had reasonable bias and MSE and acceptable 95% CI coverage. However, this measure may not always be the optimal choice for

researchers because it places equal weight on sensitivity and specificity. Summary measures that equally weight the individual components of diagnostic accuracy have been criticized (Nietert et al., 2007). As well, Youden's (1950) index can assume the same value for very different estimates of sensitivity and specificity. Youden's (1950) index may not be more reliable and useful than its individual component measures (Zaslavsky, Shaul, Zaborski, Cioffi, & Cleary, 2002).

This research demonstrated, via the numeric example, how to apply these models to assist researchers in selecting an optimal case definition in a validation study; inferential analyses were used to identify the case definition characteristics that were associated with the selected diagnostic validity measures. While descriptive statistics are useful for comparing diagnostic validity estimates, they do not account for error variation in the estimates. Moreover, it is difficult to compare multiple diagnostic validity estimates (Widdifield, Bernatsky, et al., 2013). Using the univariate or bivariate models allows for a larger number of case definitions to be tested. The accuracy in the model parameter estimation increases as the number of case definitions increase. Overall, modeling the case definitions allows researchers to justify their decision of the optimal case definition(s).

5.3 Implications

The univariate models of sensitivity and specificity performed better than the bivariate model of sensitivity and specificity. The bivariate model may result in an excessive rate of Type 1 errors and lack model convergence. However, if sample size is sufficiently large, the bivariate model can assist researchers to better understand case definition characteristics that jointly include sensitivity and specificity (Reitsma et al., 2005).

The results of a diagnostic validation study are typically used to recommend one or more case definitions that will result in maximum case ascertainment accuracy. Separate models were developed for all age groups and for the 65+ age group based on the case definition characteristics that were associated with sensitivity and specificity. Based on the univariate models of sensitivity and specificity, the modeling results suggest that the optimal case definition has 2+ physician diagnoses with 1+ physician diagnosis from a specialist and uses an unlimited observation period. The requirement of 2+ physician diagnoses were chosen because this case definition characteristic resulted in a statistically significant increase in sensitivity and small decrease in specificity compared to using the criteria of 1+ or 3+ physician diagnoses. An unlimited observation time was associated with a statistically significance increase in sensitivity and no statistically significant change in specificity relative to a 1 year observation time. Compared to no specialist diagnoses, 1+ specialist diagnosis was associated with a statistically significant increase in specificity and no statistically significant change in sensitivity. A time separation of at least 60 days was not used as it was significantly associated with a decrease in sensitivity and no change in specificity. However, a case definition with all of the significantly associated characteristics was not tested in the validation study but a second option was to use 1 year of observation time because the other observation times were not associated with either sensitivity or specificity and provides a simpler case definition. This resulted in the optimal case definition of 2+ physician diagnoses in 1 year with at least 1 physician diagnosis from a specialist in the 20+ population with sensitivity of 0.98, specificity of 0.83, PPV of 0.73, and NPV of 0.99 (Widdifield, Bernatsky, et al., 2013). In the 65+ population this case definition resulted in a sensitivity of 0.98, specificity of 0.86, PPV of 0.80, and NPV of 0.99.

For univariate sensitivity and specificity models applied to case definition data for the 65+ population the inferential analyses suggest the optimal case definition was 2+ physician diagnoses in 1 year, with 1+ of the physician diagnoses from a specialist, and 1+ DMARD or biological prescription. At least one DMARD or biological prescription was considered in the optimal case definition since it was significantly associated with an increased specificity and was not associated with sensitivity when compared to not requiring a DMARD or biological prescription. However, this case definition was not tested in the original study but a case definition of these criteria or 1+ hospital visit was tested. This case definition was considered as optimal because including 1+ hospital visit was not significantly associated with either sensitivity or specificity compared to not requiring 1+ hospital visit.

The univariate model of Youden's index and bivariate model of sensitivity and specificity displayed the same significant characteristics as the univariate models of sensitivity and specificity; thus, resulting in the same recommended case definitions. These definitions can be considered optimal if the priority is to maximize sensitivity over specificity. When prioritizing specificity or balancing sensitivity and specificity, the recommended definition would identify the case definition characteristics positively associated with specificity and Youden's (1950) index, respectively. These definitions are supported from a clinical perspective since RA diagnoses in the reference standard were mostly managed in outpatient settings; therefore, adding a diagnosis from hospital records should not improve the diagnostic validity measures. Additionally, the reference standard was collected from RA specialist records; therefore, including 1+ diagnosis by a specialist in the case definition will automatically increase the true positives and therefore increase sensitivity.

While this research focused on methods to analyze case definition data from a diagnostic validation study, an equally important consideration is the design of the validation study. To ensure that the effects of the case definition characteristics can be tested, we recommend validation studies should be designed with consideration of the principles of experimental design, such as crossing or nesting of factors. A factorial experimental design has all characteristics crossed (i.e., all characteristic combinations are tested) (Montgomery, 2012; Chapter 1). A nested experimental design consists of characteristics that can only be used within a certain characteristic category (Montgomery, 2012; Chapter 1). For example, the number of prescriptions was a nested characteristic in the RA diagnostic validation study. By ensuring the characteristics are completely crossed or nested allows for more accurate modeling estimation.

5.3.1 General Application to Model Diagnostic Validity Estimates

Figure 5.1 provides a visual tool to assist researchers who wish to model the results of a diagnostic validation study. Four steps in the modeling process are defined for the researcher to follow: 1) select diagnostic validity measures, 2) choose a model, 3) identify case definition characteristics to include in the model, and 4) interpret the estimates.

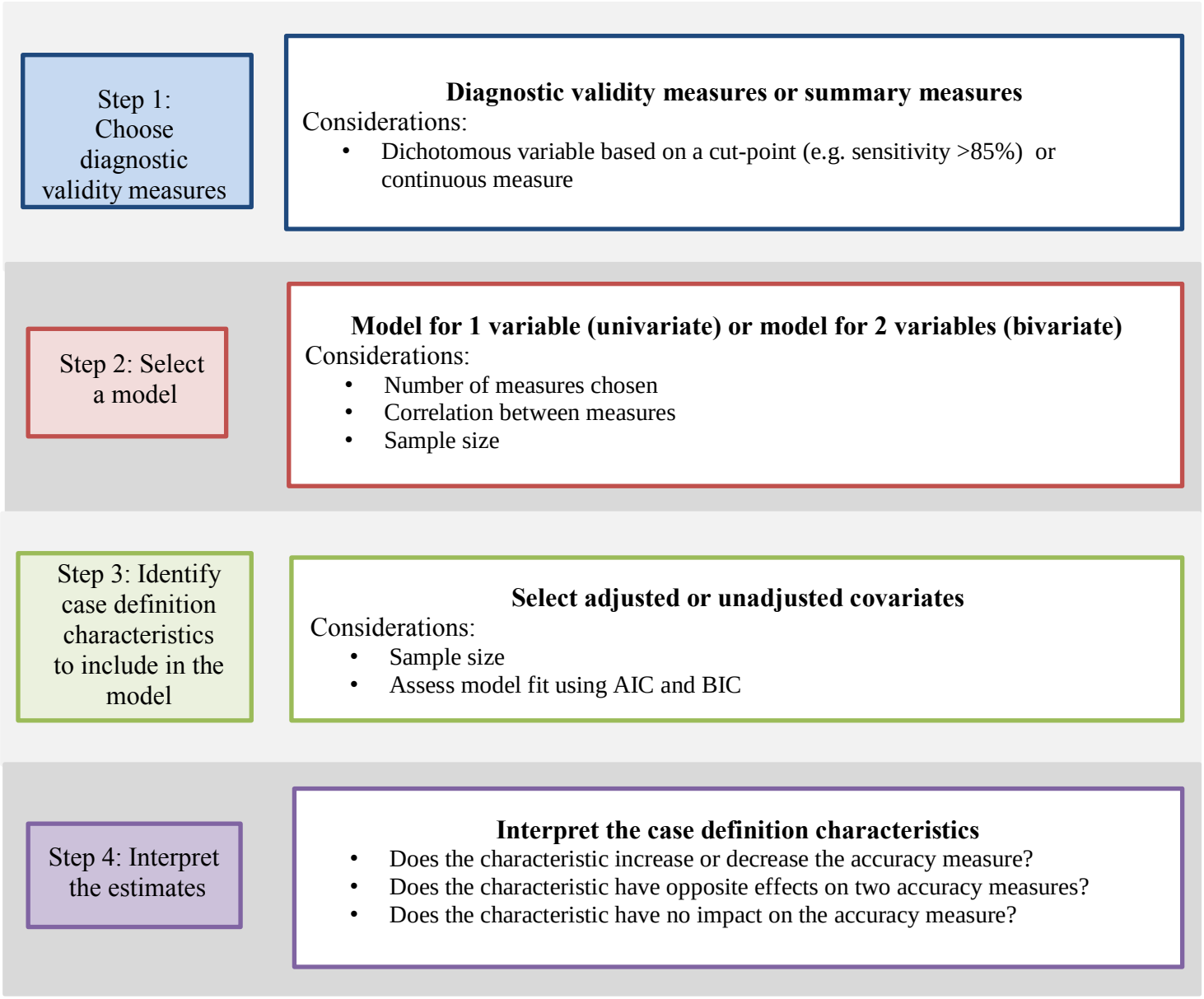
In the first step, the researcher identifies the diagnostic validity measure(s) they are most interested in using to assess case definition accuracy. Benchimol et al. (2011) suggests that a validations study should report at least four diagnostic accuracy measures. If validation studies comply, there will be four measures to choose from for modeling which allows for many model possibilities. The researcher must decide which measures are the most important in identifying the optimal case definition and the relationship between the measures chosen. A consideration at this step is whether the researcher is interested in modeling a continuous measure or a dichotomous variable based on a cut-point.

In the second step, the researcher decides on a univariate or bivariate model. A univariate model should be selected if only one diagnostic validity measure was chosen in the first step; however, if two or more measures are chosen, a bivariate model may be considered. The bivariate model is recommended when the sample size is large.

The third step requires the selection of case definition characteristics for the model. The characteristics can be chosen to be fit as unadjusted or adjusted covariates depending on sample size. As a rule of thumb, there should be approximately 10 case definitions for every parameter to be estimated (Harrell, Lee, Matchar, & Reichert, 1985). Therefore, a small sample size may limit the number of case definition characteristics that can be tested in each model. Once the model has been fit, the AIC and BIC should be used to compare nested models.

The last step focuses on interpretation of model estimates. It is important to describe if the statistically significant case definition characteristic result in an increase or decrease of the diagnostic validity measure. If the characteristic shows an increase in the measure then it can be suggested as a characteristic to be included in the optimal case definition. If the characteristic decreases the measure then the characteristic should not be included in the optimal case definition. However, if a case definition characteristic increases one measure but decreases another measure, this suggests that both options are optimal case definitions depending on the use of the case definition. Lastly, the non-significant case definition characteristics suggest that the characteristic should not be used in the optimal definition. This shows that the characteristic does not impact the diagnostic validity measure in a significant way and when taken out of the case definition it can simplify the data requirements.

Figure 5.1: Steps to model diagnostic validation estimates



By using these models instead of relying on descriptive statistics, it allows researchers to simplify the case definitions by identifying the characteristics that are not significantly associated with the diagnostic validity measures. It is important to remember that a validation study produces estimates of diagnostic validity and these models formally test the relationship between the diagnostic validity measure and the case definition characteristics.

This research contributes to improved methods for using administrative health databases to study chronic diseases. The study benefits researchers who conduct diagnostic validation studies to identify optimal case definition(s) by allowing them to test hypotheses about the case definition characteristics that influence measures of diagnostic validity. Researchers can use a model-based approach to make empirical decisions when choosing and applying a case definition from a diagnostic validation study. Applying accurate case definitions allows for unbiased population estimates of disease prevalence, which contributes to better public health decision making. Unbiased study models can be used in future research to predict values of sensitivity and specificity from case definitions.

This research improves the methods for identifying the optimal RA case definition. Therefore, this research ultimately leads to a better understanding of the impact of RA on the healthcare system, and improvements in disease treatment and management. The Public Health Agency of Canada and the Canadian Institute for Health Information are key potential users of these methods as they develop disease surveillance tools and reports (Public Health Agency of Canada, 2009, 2010)

5.4 Strengths and Limitations

This study was not without limitations. Firstly, the simulation study only included a limited number of scenarios. Not all combinations of means, variances, and correlations of sensitivity and specificity could be examined; however, the use of simulation parameter values derived from existing RA validation studies ensured that the results of the simulation study were relevant to real studies.

Secondly, the selected models were not the only ones that could be applied to the data. Other parametric as well as non-parametric models could be used such as, a bivariate random-

effects model with copulas techniques to model the correlation between sensitivity and specificity (Kuss et al., 2014). The models applied in this study were considered to be the most familiar to researchers.

The first strength of the study was the number of simulations used was based the following three criteria: (a) the number of simulations used in other similar simulation studies, (b) calculation of the number of simulations, and (c) computer simulation completion time. Previous simulation studies that applied models to diagnostic validation studies conducted 1000 simulations of each scenario (Kuss et al., 2014; Riley et al., 2007). Burton et al. (2006) proposed a simple calculation to determine the number of simulations required to achieve a specified level of accuracy:

$$B = \left(\frac{Z_{1-(\alpha/2)}}{\delta}\right)^2, \tag{5.2}$$

where $Z_{1-(\alpha/2)}$ is the $1-(\alpha/2)$ quantile of the standard normal distribution, σ is the standard deviation of the parameter of interest, δ is the acceptable level of accuracy of the parameter, and B is the number of simulations. The number of simulations needed to achieve 1% accuracy in the bias with a 5% significance level and a bias variance of 0.03 is 1,153. Using 1000 and 2000 simulations, the computer completion time was approximately 1.5 hours and 4 hours, respectively. The decision to use 2000 simulations produced a balance amongst these three criteria and ensured that an acceptable level of accuracy in the parameters was met.

A second strength of this study was the use of both computer simulation and a real numeric example. The computer simulation was used to study the performance of the statistical models for known population characteristics and the numeric example demonstrated the application of the simulation models in the real world.

A third strength was the univariate model results from this simulation study can be generalized to models of PPV and NPV, or to models for other summary diagnostic validity measures. Models of PPV and NPV will have similar performance to the results of sensitivity and specificity. The bivariate model results can also be generalized to other joint estimates (e.g., sensitivity, specificity, PPV, NPV, and/or summary diagnostic validity measures).

Another strength of this study was the assumption of a beta distribution for simulating sensitivity, specificity, and Youden's (1950) index. Many published meta-analysis diagnostic validation studies have assumed a normal distribution for sensitivity and specificity; however, the beta distribution has recently been shown to provide a better fit to the data (Hoyer & Kuss, 2015; Kuss et al., 2014). Fifth, the simulation parameters were based on a summary of real RA validation studies. Lastly, the author of the numeric example provided all case definitions tested to eliminate the potential for publication bias.

5.5 Future Research

This research has many possible extensions and directions. First, the methodology can be applied to other measures, such as a weighted Youden's (1950) index, the DOR and the TAP index. Secondly, the methodology could be extended to model case definitions within and between studies. As well, this research could be applied to validation studies for other chronic diseases.

Lastly, research is needed to provide guidance and recommendations around designing a validation study. Currently, researchers test the case definitions they think will produce high sensitivity, specificity, PPV, and NPV; however, when using methods to test for characteristics that are associated with an optimal case definition, the combination of characteristics tested together is extremely important. This importance showed in a few different ways for the numeric

example. The first is that any characteristics that are always or mostly tested together cannot be included in the same model due to multicollinearity; therefore, the effect of those characteristics cannot be determined individually. The second is any characteristic that is mainly present or absent in case definitions cannot be included in the models as they do not provide enough power for estimation of the parameter. Designing a validation study with completely crossed case definition characteristics provides the highest power and the best ability to detect the case definition characteristics that are associated with an optimal case definition.

REFERENCES

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Benchimol, E. I., Manuel, D. G., To, T., Griffiths, A. M., Rabeneck, L., & Guttmann, A. (2011). Development and use of reporting guidelines for assessing the quality of validation studies of health administrative data. *Journal of Clinical Epidemiology*, *64*(8), 821–829.

Bernatsky, S., Dekis, A., Hudson, M., Pineau, C., Boire, G., Fortin, P., Bessette, L., Jean, S., Chetaille, A.-L., Belisle, P., Bergeron, L., Feldman, D., & Joseph, L. (2014). Rheumatoid arthritis prevalence in Quebec. *BMC Research Notes*, *7*(1), 937.

Bernatsky, S., Lix, L., O'Donnell, S., Lacaille, D., & CANRAD Network. (2013). Consensus statements for the use of administrative health data in rheumatic disease research and surveillance. *Journal of Rheumatology*, *40*(1), 66–73.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. a, Glasziou, P. P., Irwig, L. M., Lijmer, J. G., Moher, D., Rennie, D., & de Vet, H. C. W. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *Radiology*, *226*(1), 24–28.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*, 4279–4292.

Campbell, H., Biloglav, Z., & Rudan, I. (2008). Reducing Bias from Test Misclassificiation in Burden of Disease Studies: Use of Test to Actual Positive Ratio - New Test Parameter. *Croat Med Journal*, *49*, 402–14.

Carrara, G., Scire, C. A., Zambon, A., Cimmino, M. A., Cerra, C., Caprioli, M., Cagnotto, G., Nicotra, F., Arfe, A., Migliazza, S., Corrao, G., Minisola, G., & Montecucco, C. (2015). A validation study of a new classification algorithm to identify rheumatoid arthritis using administrative health databases: case-control and cohort diagnostic accuracy studies. *BMJ Open*, *5*(1), e006029.

Chen, F., Xue, Y., Tan, M. T., & Chen, P. (2015). Efficient statistical tests to compare Youden index: Accounting for contingency correlation. *Statistics in Medicine*, *34*, 1560–1576.

Chu, H., Guo, H., & Zhou, Y. (2010). Bivariate random effects meta-analysis of diagnostic studies using generalized linear mixed models. *Medical Decision Making*, *30*(4), 499–508.

Chu, H., Nie, L., Cole, S. R., & Poole, C. (2009). Meta-analysis of diagnostic accuracy studies accounting for disease prevalence: Alternative parameterizations and model selection. *Statistics in Medicine*, *28*, 2384–2399.

Chubak, J., Pocobelli, G., & Weiss, N. S. (2012). Trade-offs between accuracy measures for electronic healthcare data algorithms. *Journal of Clinical Epidemiology*, *65*(3), 343–9.

Cleophas, T. J., & Zwinderman, A. H. (2009). Meta-analyses of diagnostic studies. *Clinical Chemistry and Laboratory Medicine*, *47*(11), 1351–1354.

Coombs, W. T., & Algina, J. (1996). On Sample Size Requirements for Johansen ' s Test. *Journal of Educational and Behavioral Statistics*, *21*(2), 169–178.

Dart, A. B., Martens, P. J., Sellers, E. A., Brownell, M. D., Rigatto, C., & Dean, H. J. (2011). Validation of a pediatric diabetes case definition using administrative health data in Manitoba, Canada. *Diabetes Care*, *34*(4), 898–903.

Diaz, M. (2015). Performance measures of the bivariate random effects model for meta-analyses of diagnostic accuracy. *Computational Statistics & Data Analysis*, *83*, 82–90.

Emrich, L. J., & Piedmonte, M. R. (1991). A Method for Generating High-Dimensional Multivariate Binary Variables. *The American Statistician*, *45*, 302–304.

Fung, V., Brand, R. J., Newhouse, J. P., & Hsu, J. (2011). Using Medicare data for comparative effectiveness research: opportunities and challenges. *The American Journal of Managed Care*, *17*(7), 488–496.

Gabriel, S. E. (1994). The sensitivity and specificity of computerized databases for the diagnosis of rheumatoid arthritis. *Arthritis*, *37*(6), 821–823.

Giavarina, D. (2012). Tools for critical appraisal of evidence in studies of diagnostic accuracy. *Autoimmunity Reviews*, *12*(2), 89–96.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology*, *56*(11), 1129–1135.

Harbord, R. M., Whiting, P., Sterne, J. A. C., Egger, M., Deeks, J. J., Shang, A., & Bachmann, L. M. (2008). An empirical comparison of methods for meta-analysis of diagnostic accuracy showed hierarchical models are necessary. *Journal of Clinical Epidemiology*, *61*(11), 1095–1103.

Harrell, F., Lee, K., Matchar, D., & Reichert, T. (1985). Regression models for prognostic prediction: advantages, problems, and suggested solutions. *Cancer Treat Rep*, *69*(10), 1071–77.

Harrold, L. R., Salman, C., Shoor, S., Curtis, J. R., Asgari, M. M., Gelfand, J. M., Wu, J. J., & Herrinton, L. J. (2013). Incidence and prevalence of juvenile idiopathic arthritis among children in a managed care population, 1996-2009. *The Journal of Rheumatology*, *40*(7), 1218–25.

Hoyer, A., & Kuss, O. (2015). Meta-analysis of diagnostic tests accounting for disease prevalence: a new model using trivariate copulas. *Statistics in Medicine*, *34*, 1912–1924.

Hux, J. E., Ivis, F., Flintoft, V., & Bica, A. (2002). Diabetes in Ontario: determination of prevalence and incidence using a validated administrative data algorithm. *Diabetes Care*, *25*(3), 512–516.

Jolley, R. J., Sawka, K. J., Yergens, D. W., Quan, H., Jetté, N., & Doig, C. J. (2015). Validity of administrative data in recording sepsis: a systematic review. *Critical Care*, *19*(1), 139.

Jones, C. M., Ashrafian, H., Darzi, A., & Athanasiou, T. (2010). Guidelines for diagnostic tests and diagnostic accuracy in surgical research. *Journal of Investigative Surgery*, *23*(1), 57–65.

Katz, J. N., Barrett, J., Liang, M. H., Bacon, A. M., Kaplan, H., Kieval, R. I., Lindsey, S. M., Roberts, W. N., Sheff, D. M., Spencer, R. T., Weaver, A. L., & Baron, J. A. (1997).

Sensitivity and positive predictive value of medicare part B physician claims for rheumatologic diagnoses and procedures. *Arthritis & Rheumatism*, *40*(9), 1594–1600.

Kim, S. Y., Servi, A., Polinski, J. M., Mogun, H., Weinblatt, M. E., Katz, J. N., & Solomon, D. H. (2011). Validation of rheumatoid arthritis diagnoses in health care utilization data. *Arthritis Research & Therapy*, *13*(1), R32.

Kuss, O., Hoyer, A., & Solms, A. (2014). Meta-analysis for diagnostic accuracy studies: a new statistical model using beta-binomial distributions and bivariate copulas. *Statistics in Medicine*, *33*(1), 17–30.

Lai, C.-Y., Tian, L., & Schisterman, E. F. (2012). Exact confidence interval estimation for the Youden index and its corresponding optimal cut-point. *Computational Statistics and Data Analysis*, *56*(5), 1103–1114.

Leite, W. L., Jimenez, F., Kaya, Y., Stapleton, L. M., Macinnes, W., & Sandbach, R. (2015). An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies An Evaluation of Weighting Methods Based on Propensity Scores to Reduce Selection Bias in Multilevel Observational Studies. *Multivariate Behavioral Research*, *50*(3), 265–284.

Leslie, W. D., Lix, L. M., & Yogendran, M. S. (2011). Validation of a case definition for osteoporosis disease surveillance. *Osteoporosis International*, *22*, 37–46.

Li, D., Shen, F., Yin, Y., Peng, J., & Chen, P. (2013). Weighted Youden index and its two-independent-sample comparison based on weighted sensitivity and specificity, *126*(6), 1150–1154.

Li, J., & Fine, J. P. (2011). Assessing the dependence of sensitivity and specificity on prevalence in meta-analysis. *Biostatistics*, *12*(4), 710–722.

Lix, L. M., Yan, L., Blackburn, D., Hu, N., Schneider-Lindner, V., & Teare, G. F. (2014). Validity of the RAI-MDS for ascertaining diabetes and comorbid conditions in long-term care facility residents. *BMC Health Services Research*, *14*, 17.

Lix, L. M., 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.

Lix, L. M., Yogendran, M. S., Leslie, W. D., Shaw, S. Y., Baumgartner, R., Bowman, C., Metge, C., Gumel, A., Hux, J., & James, R. C. (2008). Using multiple data features improved the validity of osteoporosis case ascertainment from administrative databases. *Journal of Clinical Epidemiology*, *61*(12), 1250–60.

Menke, J. (2010). Bivariate random-effects meta-analysis of sensitivity and specificity with SAS PROC GLIMMIX. *Methods of Information in Medicine*, *49*(1), 54–64.

Montgomery, D. C. (2012). *Design and analysis of experiments* (eigth edit). Hoboken, NJ: John Wiley & Sons, Inc.

Ng, B., Aslam, F., Petersen, N. J., Yu, H.-J., & Suarez-Almazor, M. E. (2012). Identification of rheumatoid arthritis patients using an administrative database: a Veterans Affairs study. *Arthritis Care & Research*, *64*(10), 1490–6.

Nietert, P. J., Wessell, A. M., Jenkins, R. G., Feifer, C., Nemeth, L. S., & Ornstein, S. M. (2007). Using a summary measure for multiple quality indicators in primary care: the Summary QUality InDex (SQUID). *Implementation Science*, *2*(11).

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cut-points using two ROC based criteria. *Am J Epidemiol*, *163*(7), 670–675.

Public Health Agency of Canada. (2009). *Report from the National Diabetes Surveillance System: Diabetes in Canada, 2009*. Ottawa: Public Health Agency of Canada. Retrieved from http://www.phac-aspc.gc.ca/publicat/2009/ndssdic-snsddac-09/pdf/report-2009-eng.pdf

Public Health Agency of Canada. (2010). *Report from the Canadian Chronic Disease Surveillance System: Hypertension in Canada, 2010*. Ottawa: Public Health Agency of Canada. Retrieved from http://www.phac-aspc.gc.ca/cd-mc/cvd-mcv/ccdss-snsmc-2010/pdf/CCDSS_HTN_Report_FINAL_EN_20100513.pdf

Quan, H., Khan, N., Hemmelgarn, B. R., Tu, K., Chen, G., Campbell, N., Hill, M. D., Ghali, W. a, & McAlister, F. a. (2009). Validation of a case definition to define hypertension using administrative data. *Hypertension*, *54*(6), 1423–8.

Reitsma, J. B., Glas, A. S., Rutjes, A. W. S., Scholten, R. J. P. M., Bossuyt, P. M., & Zwinderman, A. H. (2005). Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *Journal of Clinical Epidemiology*, *58*(10), 982–990.

Riley, R. D., Abrams, K. R., Sutton, A. J., Lambert, P. C., & Thompson, J. R. (2007). Bivariate random-effects meta-analysis and the estimation of between-study correlation. *BMC Medical Research Methodology*, *7*, 3.

SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Cary, NC.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177.

Schisterman, E. F., & Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. *Communications in Statistics - Simulation and Computation*, *36*(3), 549–563.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Shaffer, J. P. (1995). Multiple hypothesis testing. *Annu. Rev. Psychol*, *46*, 561–584.

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*, 1–15.

Singh, J. A., Holmgren, A. R., & Noorbaloochi, S. (2004). Accuracy of Veterans administration databases for a diagnosis of rheumatoid arthritis. *Arthritis and Rheumatism*, *51*(6), 952–7.

Sinharay, S., Stern, H. S., & Russel, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods*, *6*(4), 317–329.

Stringer, E., & Bernatsky, S. (2015). Validity of juvenile idiopathic arthritis diagnoses using

administrative health data. *Rheumatology International*, *35*(3), 575–9.

Suissa, S., Henry, D., Caetano, P., Dormuth, C. R., Ernst, P., Hemmelgarn, B., Lelorier, J., Levy, A., Martens, P. J., Paterson, J. M., Platt, R. W., Sketris, I., Teare, G., & Network, C. (2012). CNODES : the Canadian Network for Observational Drug Effect Studies. *Open Med*, *6*(4), 134–140.

Tang, L., Song, J., Belin, T. R., & Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine*, *24*(14), 2111–2128.

Tennis, P., Bombardier, C., Malcolm, E., & Downey, W. (1993). Validity of rheumatoid arthritis diagnoses listed in the Saskatchewan hospital separations database. *Journal of Clinical Epidemiology*, *46*(7), 675–683.

Tu, K., Campbell, N. R. C., Chen, Z.-L., Cauch-Dudek, K. J., & McAlister, F. A. (2007). Accuracy of administrative databases in identifying patients with hypertension. *Open Med*, *1*(1), E18–E26.

Vecchio, T. J. (1966). Predictive value of a single diagnostic test in unselected populations. *New England Journal of Medicine*, *274*, 1171–1173.

Weiss, N. A., Holmes, P. T., & Hardy, M. (2005). *A Course in Probability* (2nd ed.). Boston: Pearson Addison Wesley.

Whiting, P., Rutjes, A. W. S., Reitsma, J. B., Bossuyt, P. M. M., & Kleijnen, J. (2003). The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, *3*, 25.

Wicklin, R. (2013). *Simulating Data with SAS*. Cary, North Carolina: SAS Institute Inc.

Widdifield, J., Bernatsky, S., Paterson, J. M., Tu, K., Ng, R., Thorne, J. C., Pope, J. E., & Bombardier, C. (2013). Accuracy of Canadian health administrative databases in identifying patients with rheumatoid arthritis: a validation study using the medical records of rheumatologists. *Arthritis Care & Research*, *65*(10), 1582–91.

Widdifield, J., Bombardier, C., Bernatsky, S., Paterson, J. M., Green, D., Young, J., Ivers, N., Butt, D. A., Jaakkimainen, R. L., Thorne, J. C., & Tu, K. (2014). An administrative data validation study of the accuracy of algorithms for identifying rheumatoid arthritis: the influence of the reference standard on algorithm performance. *BMC Musculoskeletal Disorders*, *15*, 216.

Widdifield, J., Labrecque, J., Lix, L., Paterson, J. M., Bernatsky, S., Tu, K., Ivers, N., & Bombardier, C. (2013). Systematic review and critical appraisal of validation studies to identify rheumatic diseases in health administrative databases. *Arthritis Care & Research*, *65*(9), 1490–1503.

Youden, W. J. (1950). Index for rating diagnostic tests. *Cancer*, *3*(1), 32–35.

Zaslavsky, A. M., Shaul, J. a, Zaborski, L. B., Cioffi, M. J., & Cleary, P. D. (2002). Combining health plan performance indicators into simpler composite measures. *Health Care Financing Review*, *23*(4), 101–115.

Zhang, H., Lu, N., Feng, C., Thurston, S. W., Xia, Y., & Tu, X. M. (2011). On fitting generalized

linear mixed-effects models for binary responses using different statistical packages. *Statistics in Medicine*, *30*(20), 2562–2572.

APPENDIX A: Summary of Validation Studies and Simulation Scenarios

Table A.1: Summary of rheumatoid arthritis and juvenile rheumatoid arthritis validation studies

Study	Location	Adult or Juvenile Population	Administrative Health Data	Reference Standard	ICD-9 Codes	ICD-10 codes	Number of Case Definitions	Diagnostic Validity Measures
Tennis, 1993	Saskatchewan, Canada	Adult	Hosp	Chart review	-	-	Unknown	-
Gabriel, 1994	Rochester, USA	Adult	Hosp & phys	Chart review	-	-	1	Sensitivity, specificity, PPV, & NPV
Katz, 1997	USA	Adult	Phys	Chart review	714.0-.3 & 714.30-.33	-	1	Sensitivity & PPV
Singh, 2004	Minneapolis, USA	Adult	Hosp, phys, pharm & lab	Chart review	714	-	5	Sensitivity, specificity, PPV, & NPV
Lix, 2006	Manitoba, Canada	Adult	Hosp, phys & pharm	National survey	714	-	16	Sensitivity, specificity, PPV, & NPV
Kim, 2011	Pennsylvania, USA	Adult	Hosp, phys, pharm & lab	Chart review	714	-	18	PPV
Ng, 2012	USA	Adult	Hosp, phys & pharm	Chart review	714	-	6	Sensitivity, specificity, & PPV
Widdifield, 2013	Ontario, Canada	Adult	Hosp, phys & pharm	Chart review	714	M05 & M06	62	Sensitivity, specificity, PPV, & NPV
Harrold, 2013	California, USA	Juvenile	Hosp, phys & pharm	Chart review	696.0, 714, & 720	-	7	Sensitivity & PPV
Widdifield, 2014	Ontario, Canada	Adult	Hosp, phys & pharm	Chart review	714	M05 & M06	43	Sensitivity, specificity, PPV, & NPV
Bernatsky,	Quebec,	Adult	Hosp & phys	Modelled	714	M05	3	Sensitivity

2014	Canada							
Carrara, 2015	Italy	Adult	Hosp, phys & pharm	Chart review	714	-	19	Sensitivity & specificity
Stringer, 2015	Nova Scotia, Canada	Juvenile	Phys	Clinical database	714	-	4	Sensitivity & PPV

Note: Hosp: hospital data; Phys: physician data; Pharm: pharmacy data; Lab: laboratory data; PPV: positive predictive value; NPV: negative predictive value; ICD = International Classification of Diseases

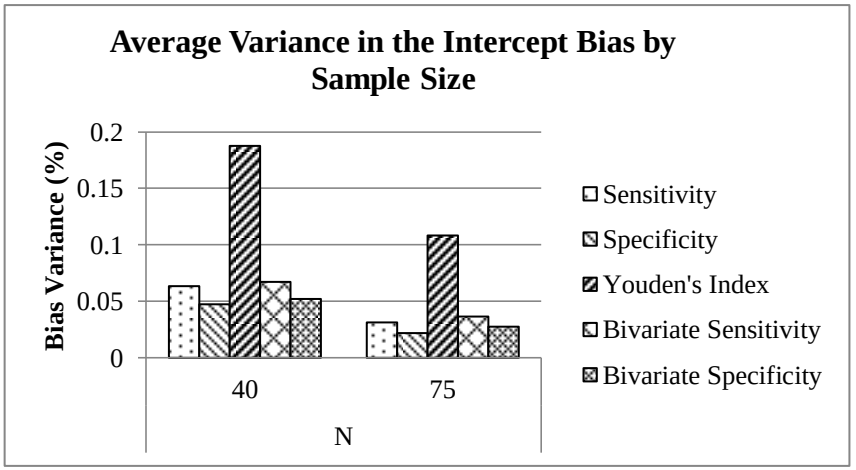
Table A.2: Simulation scenarios as defined by mean and variance of sensitivity and specificity and correlation between sensitivity and specificity

Scenario	N	μ_{sens} (σ^2_{sens})	μ_{spec} (σ^2_{spec})	$\rho_{sens\ spec}$
1	40	0.70 (0.01)	0.90 (0.01)	0
2	40	0.80 (0.01)	0.90 (0.01)	0
3	40	0.90 (0.01)	0.90 (0.01)	0
4	40	0.70 (0.03)	0.90 (0.03)	0
5	40	0.80 (0.03)	0.90 (0.03)	0
6	40	0.90 (0.03)	0.90 (0.03)	0
7	40	0.70 (0.01)	0.90 (0.01)	-0.2
8	40	0.80 (0.01)	0.90 (0.01)	-0.2
9	40	0.90 (0.01)	0.90 (0.01)	-0.2
10	40	0.70 (0.03)	0.90 (0.03)	-0.2
11	40	0.80 (0.03)	0.90 (0.03)	-0.2
12	40	0.90 (0.03)	0.90 (0.03)	-0.2
13	40	0.70 (0.01)	0.90 (0.01)	-0.7
14	40	0.80 (0.01)	0.90 (0.01)	-0.7
15	40	0.90 (0.01)	0.90 (0.01)	-0.7
16	40	0.70 (0.03)	0.90 (0.03)	-0.7
17	40	0.80 (0.03)	0.90 (0.03)	-0.7
18	40	0.90 (0.03)	0.90 (0.03)	-0.7
19	75	0.70 (0.01)	0.90 (0.01)	0
20	75	0.80 (0.01)	0.90 (0.01)	0
21	75	0.90 (0.01)	0.90 (0.01)	0
22	75	0.70 (0.03)	0.90 (0.03)	0
23	75	0.80 (0.03)	0.90 (0.03)	0
24	75	0.90 (0.03)	0.90 (0.03)	0
25	75	0.70 (0.01)	0.90 (0.01)	-0.2
26	75	0.80 (0.01)	0.90 (0.01)	-0.2
27	75	0.90 (0.01)	0.90 (0.01)	-0.2
28	75	0.70 (0.03)	0.90 (0.03)	-0.2
29	75	0.80 (0.03)	0.90 (0.03)	-0.2
30	75	0.90 (0.03)	0.90 (0.03)	-0.2
31	75	0.70 (0.01)	0.90 (0.01)	-0.7
32	75	0.80 (0.01)	0.90 (0.01)	-0.7
33	75	0.90 (0.01)	0.90 (0.01)	-0.7
34	75	0.70 (0.03)	0.90 (0.03)	-0.7
35	75	0.80 (0.03)	0.90 (0.03)	-0.7
36	75	0.90 (0.03)	0.90 (0.03)	-0.7

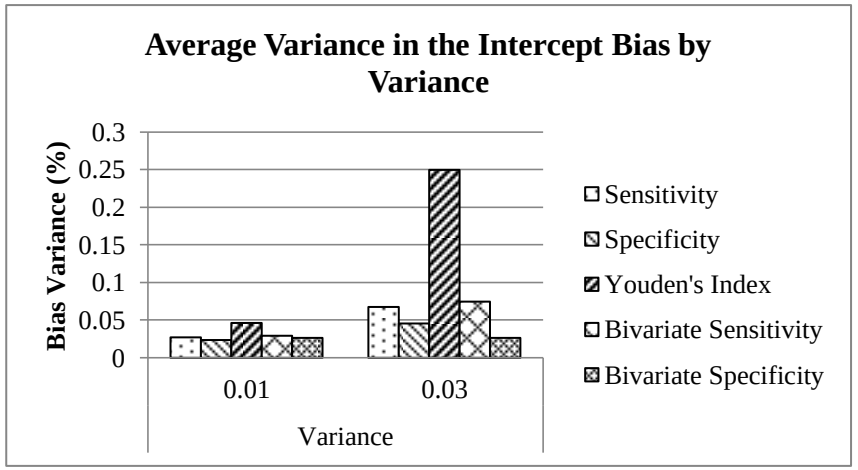
APPENDIX B: Additional Simulation Descriptive Results

Figure B.1: Average variance of the intercept bias for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

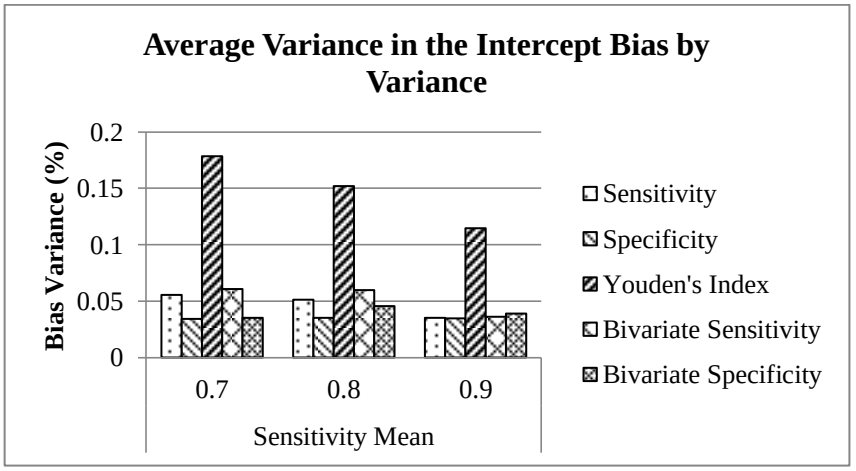
a)



b)



c)



d)

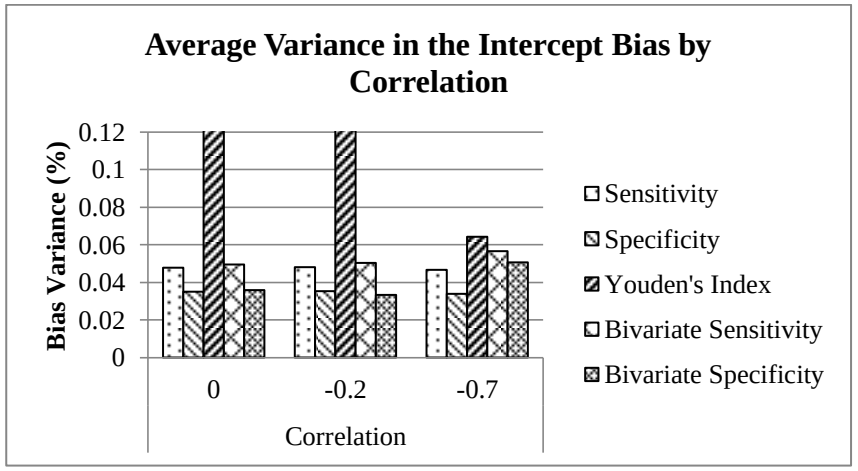
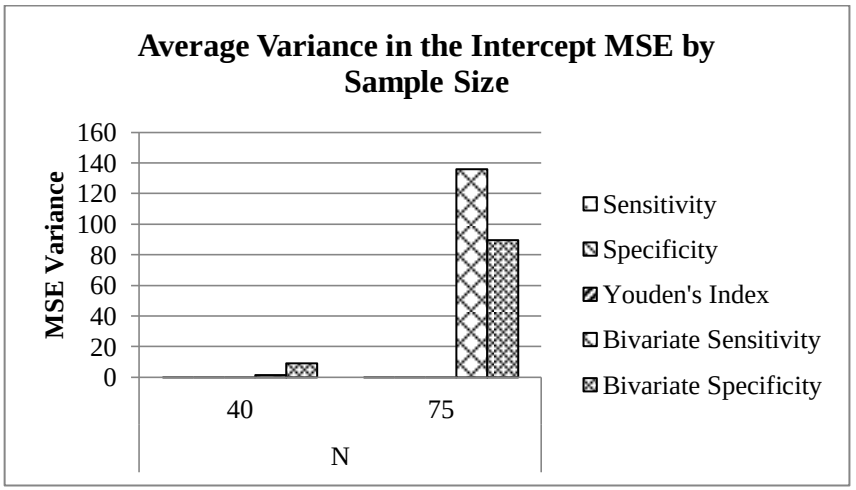
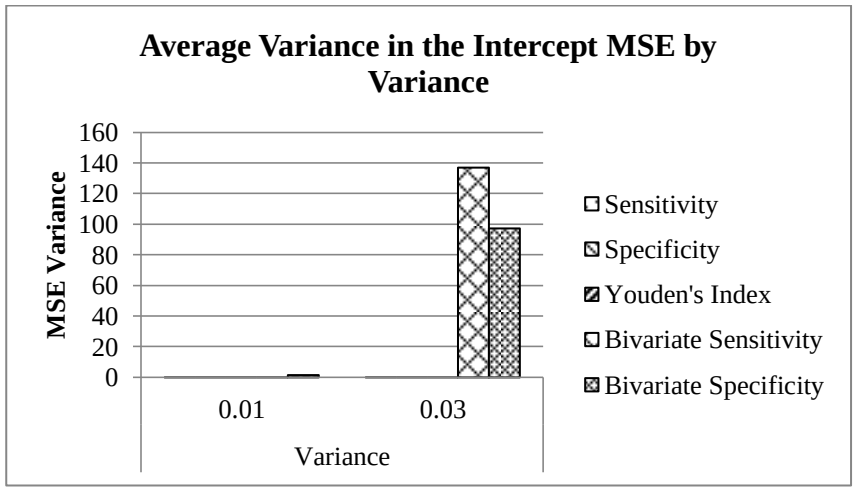


Figure B.2: Average variance of the intercept mean square error (MSE) for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

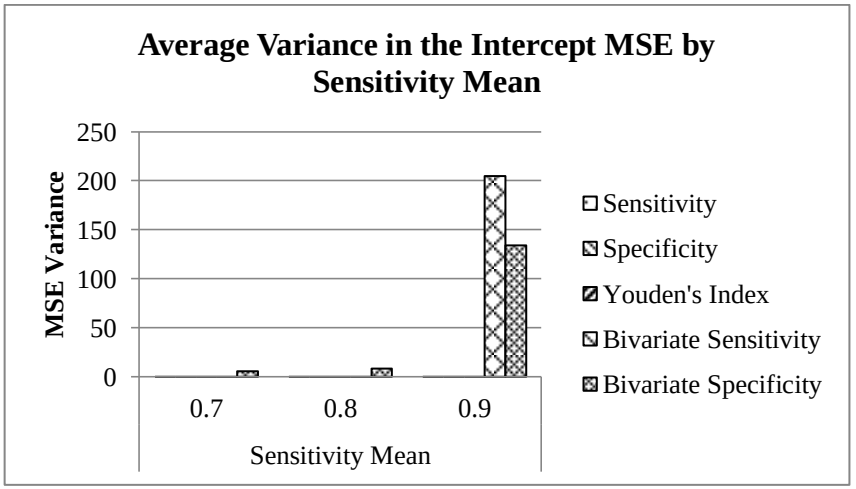
a)



b)



c)



d)

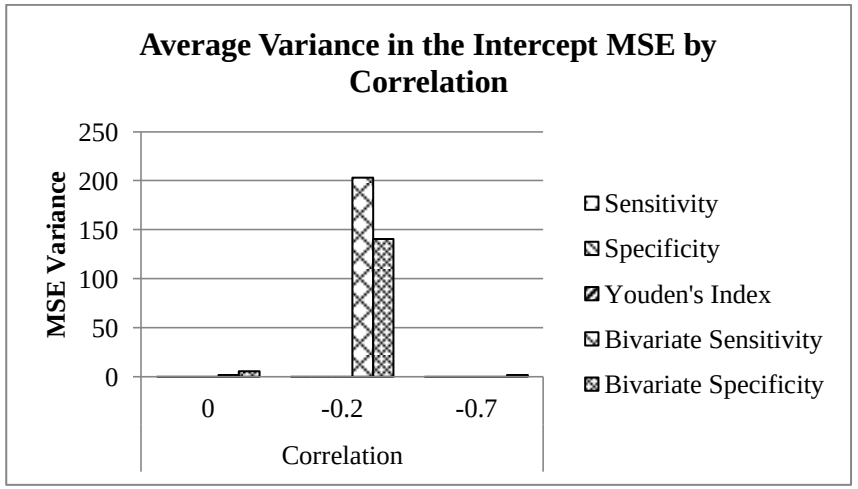
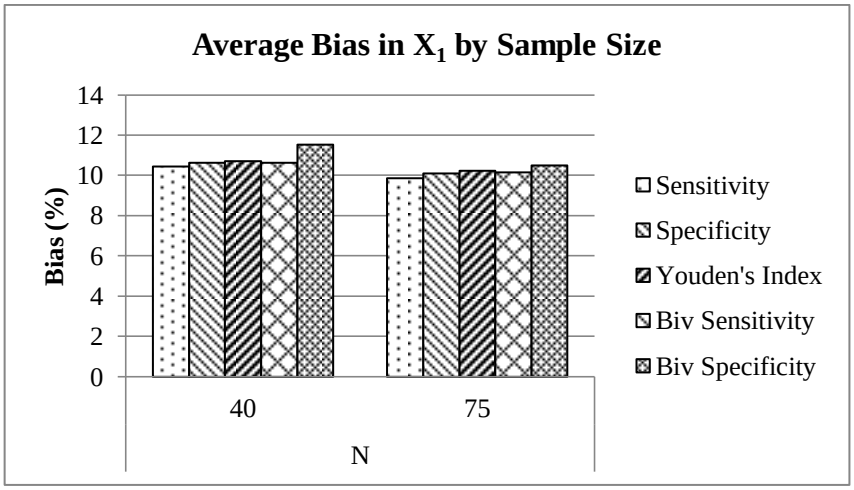
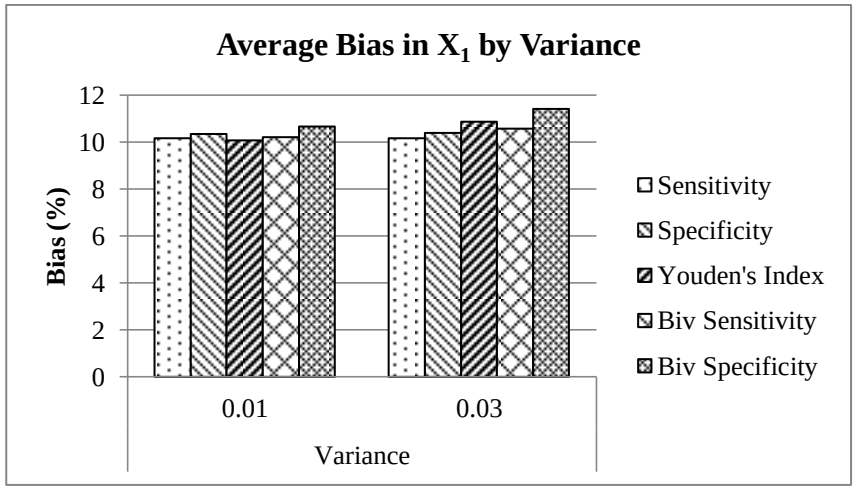


Figure B.3: Average bias in X_1 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

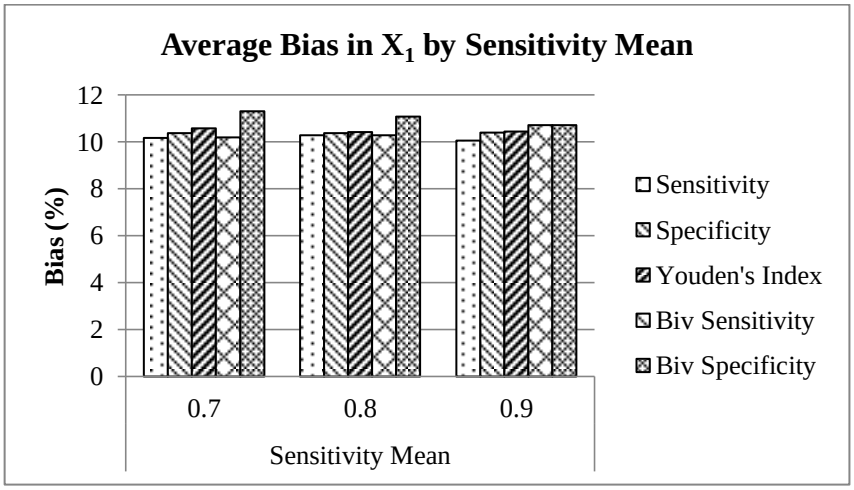
a)



b)



c)



d)

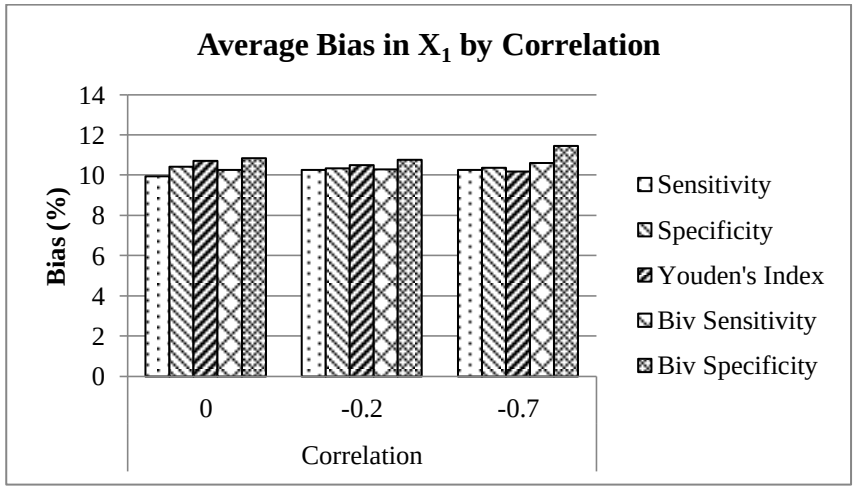
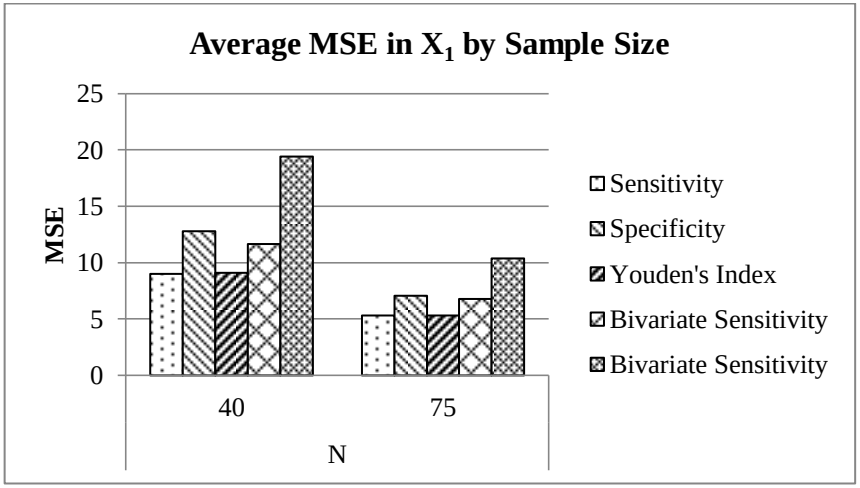
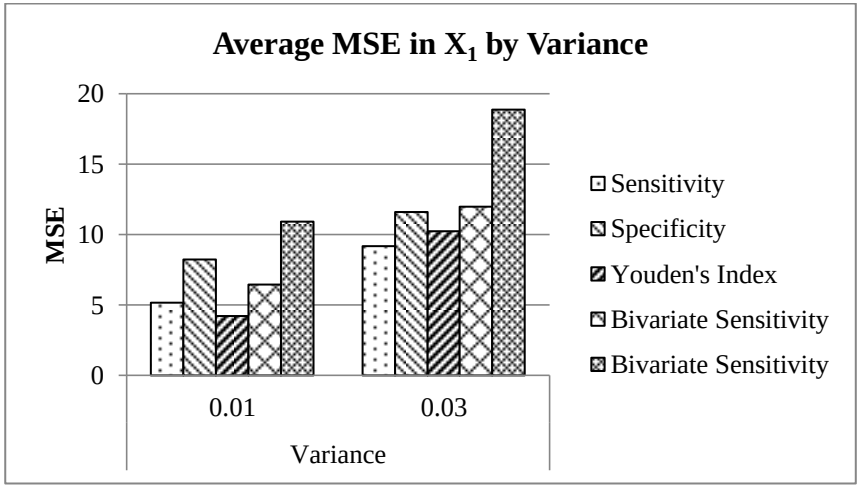


Figure B.4: Average mean square error (MSE) in X_1 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

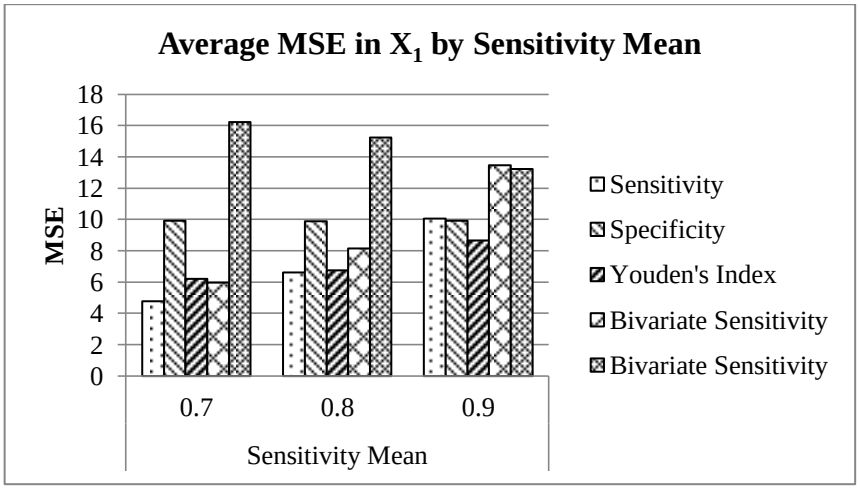
a)



b)



c)



d)

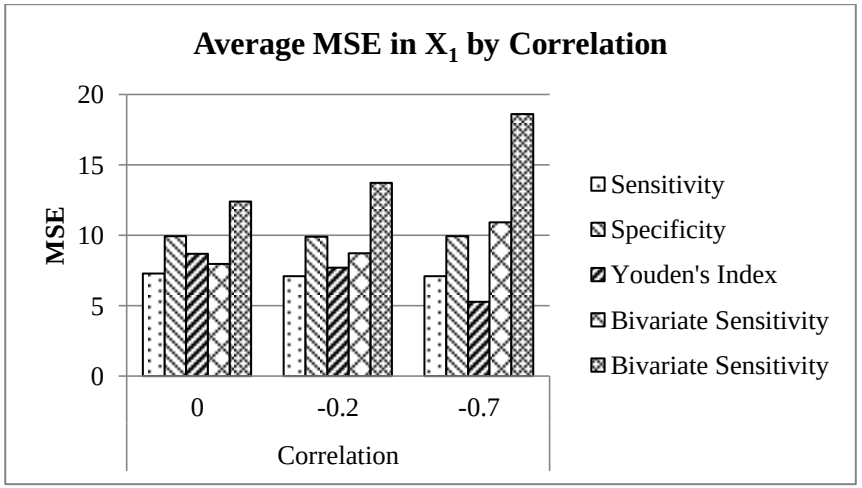
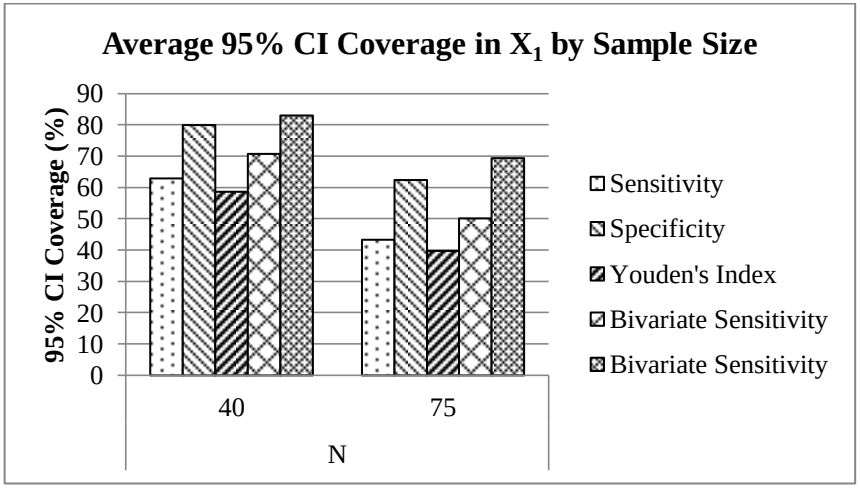
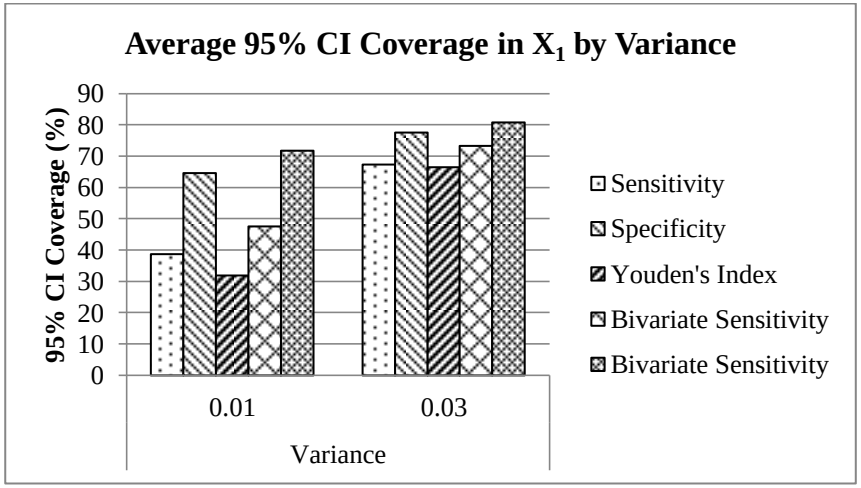


Figure B.5: Average 95% confidence interval (CI) coverage in X_1 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

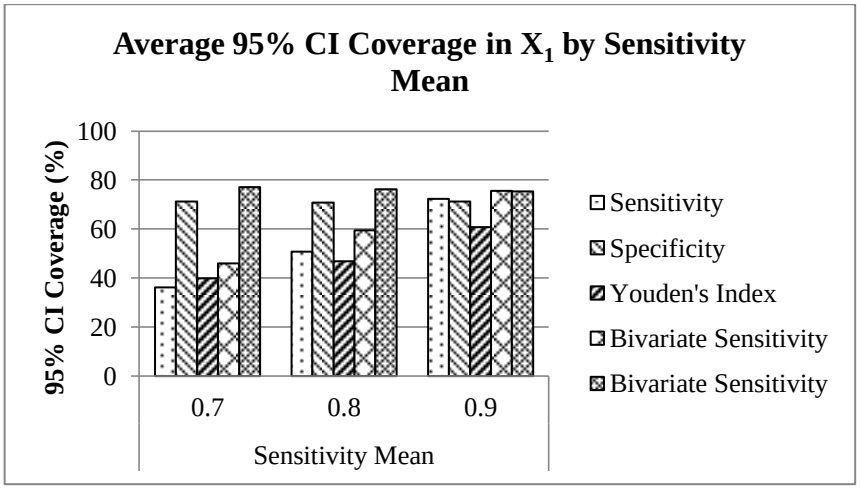
a)



b)



c)



d)

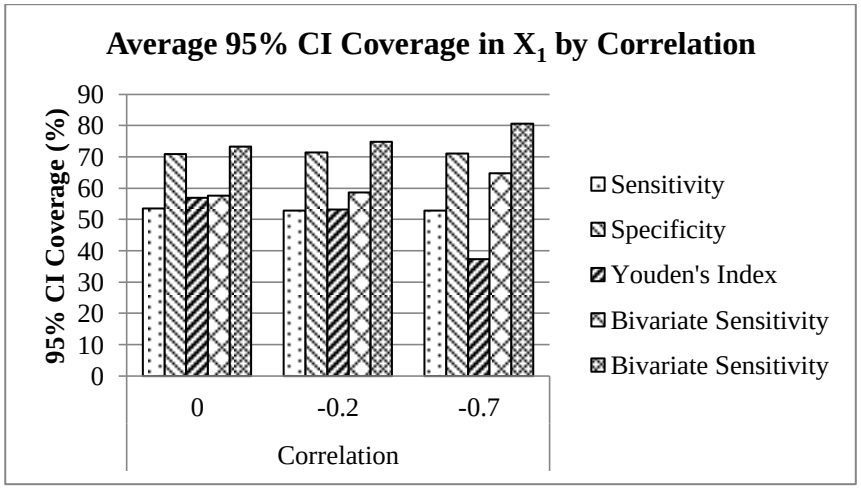
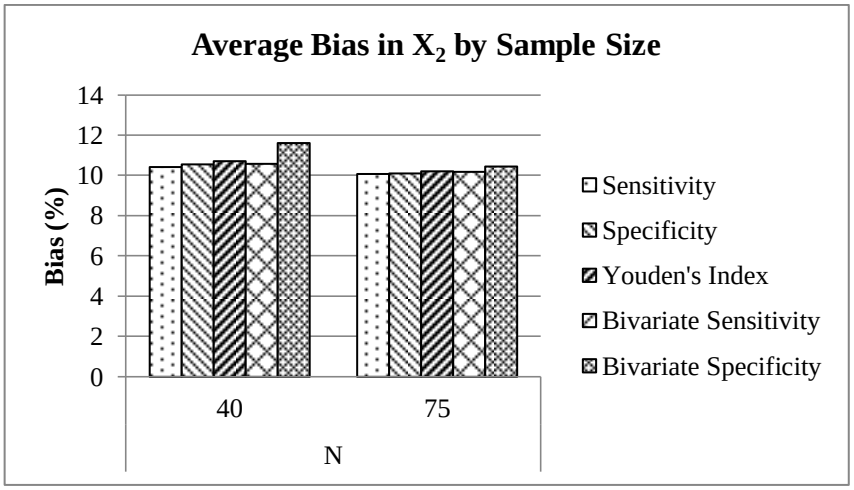
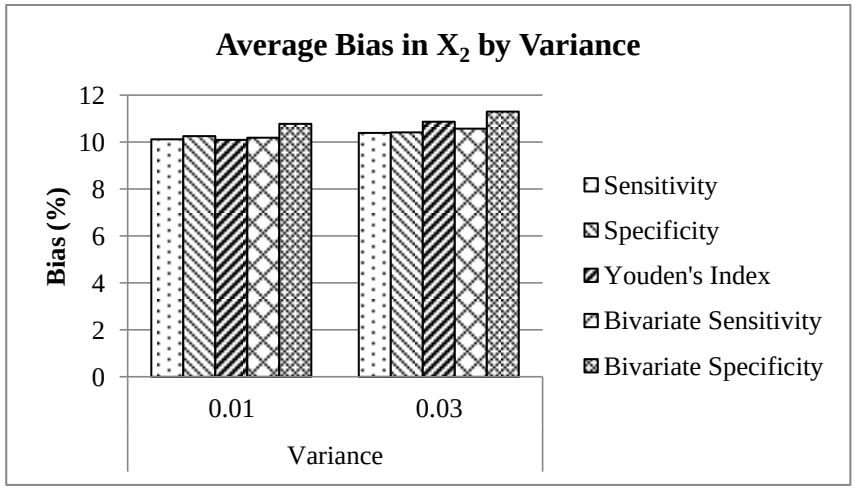


Figure B.6: Average bias in X_2 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

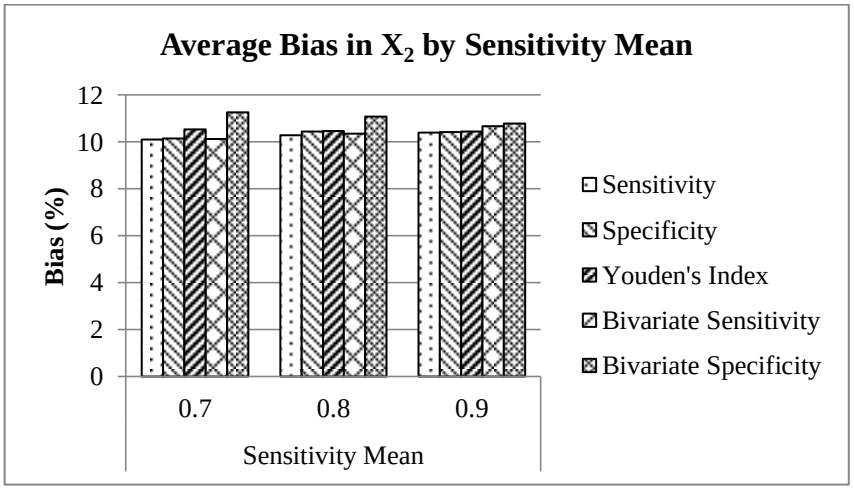
a)



b)



c)



d)

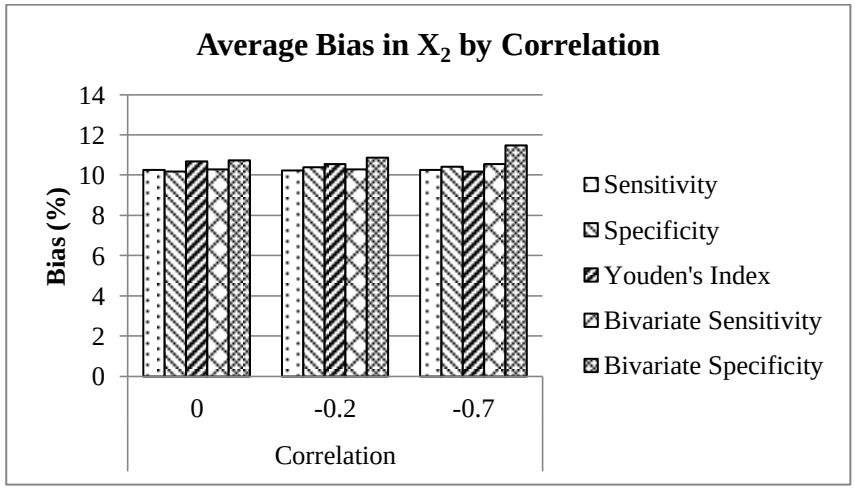
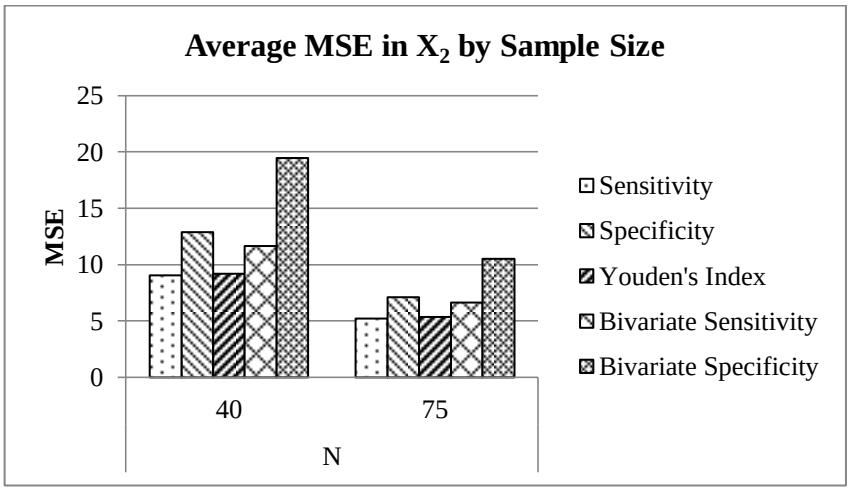
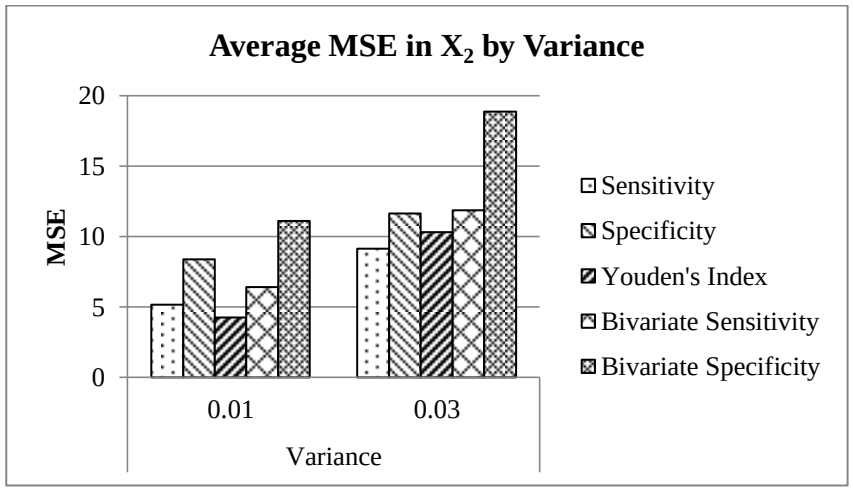


Figure B.7: Average mean square error (MSE) in X_2 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

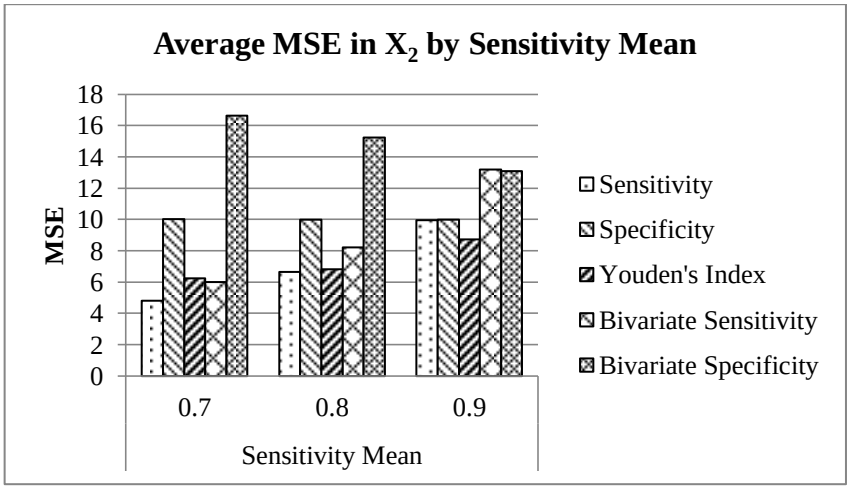
a)



b)



c)



d)

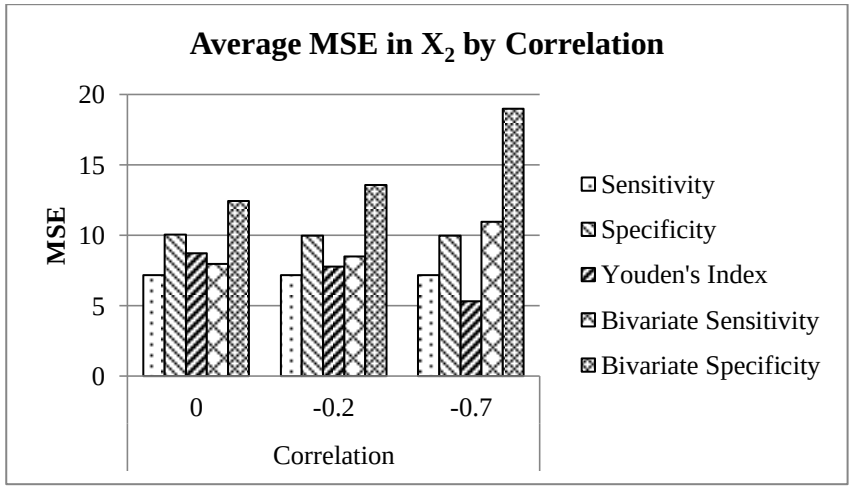
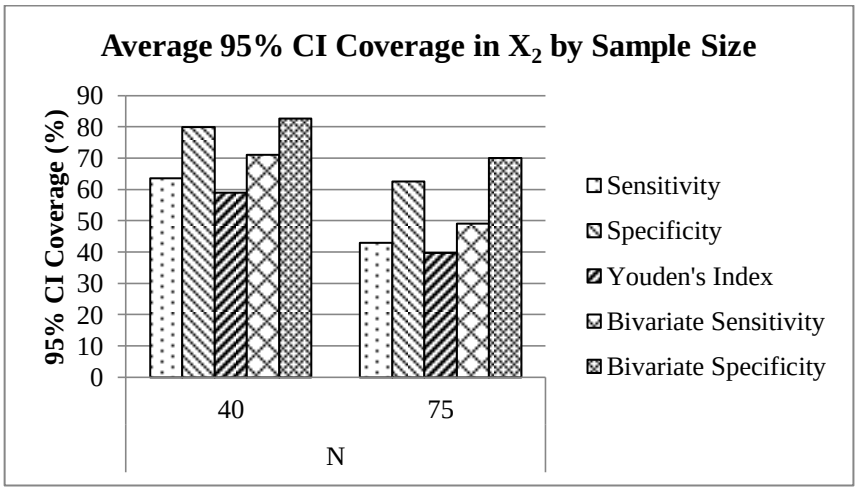
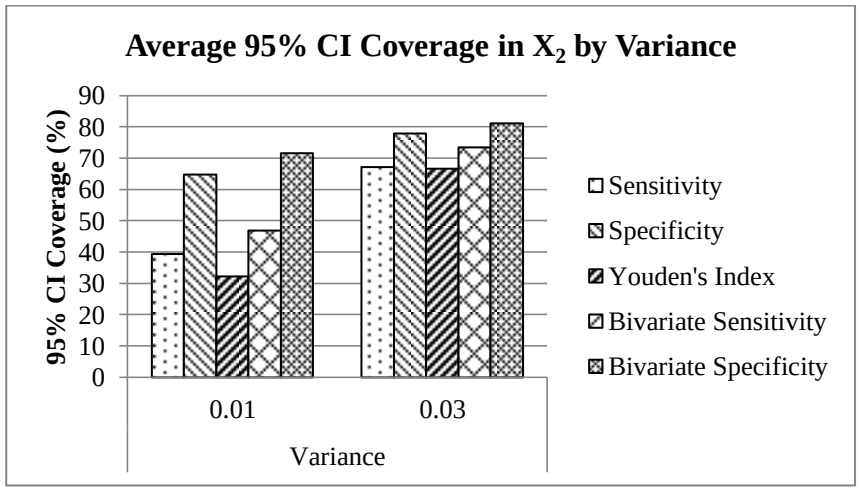


Figure B.8: Average 95% confidence interval (CI) coverage in X_2 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

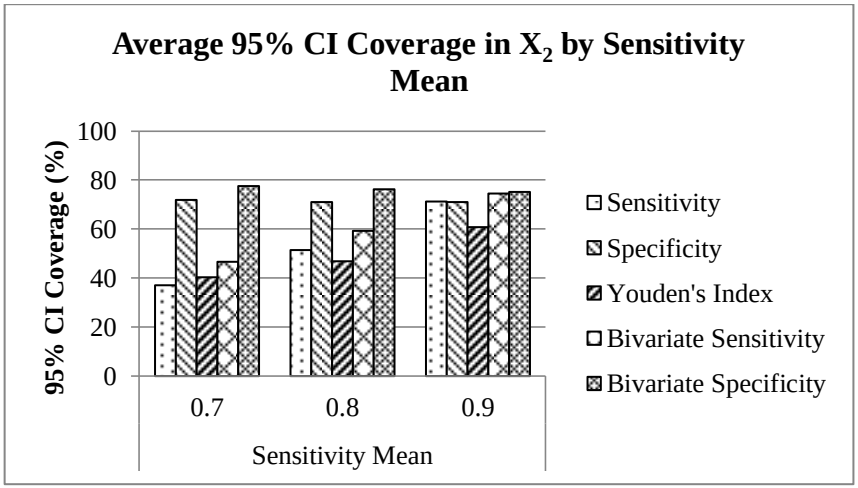
a)



b)



c)



d)

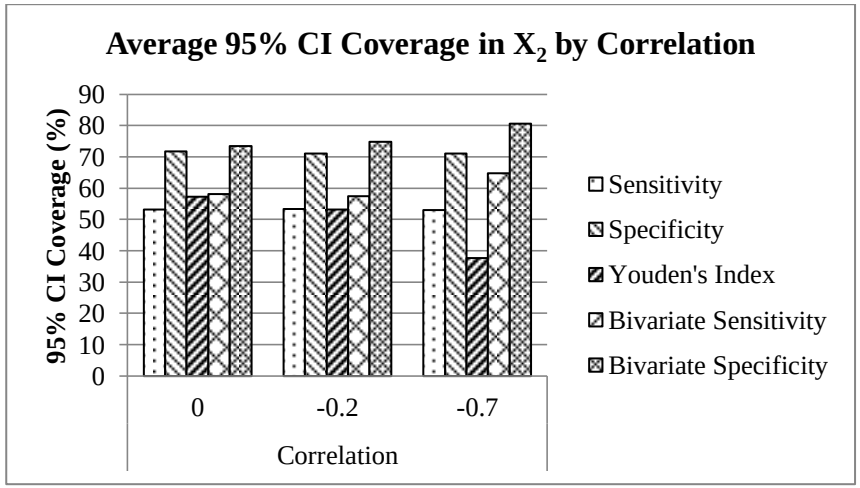
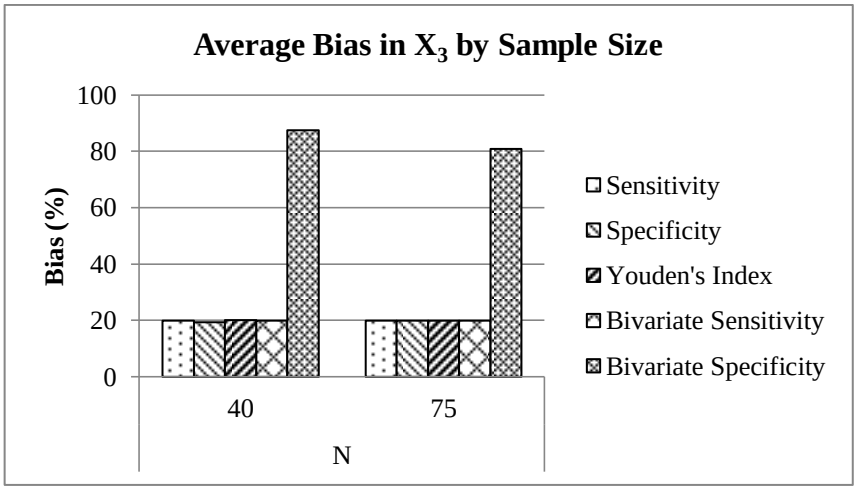
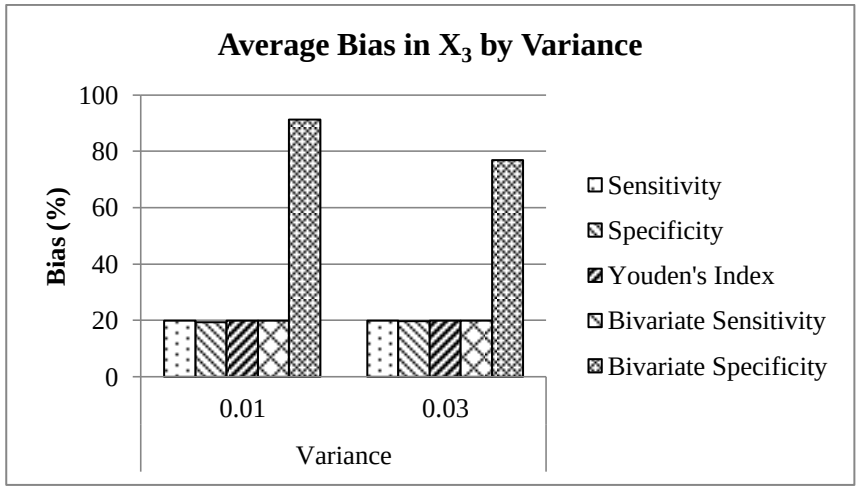


Figure B.9: Average bias in X_3 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

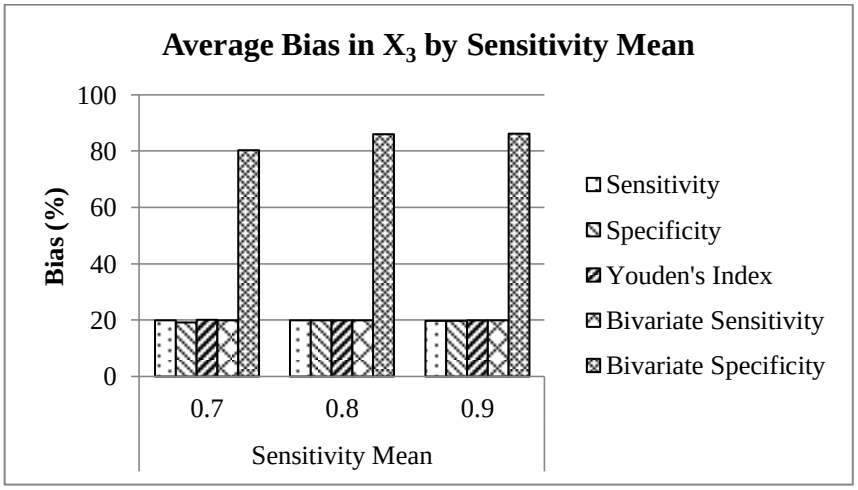
a)



b)



c)



d)

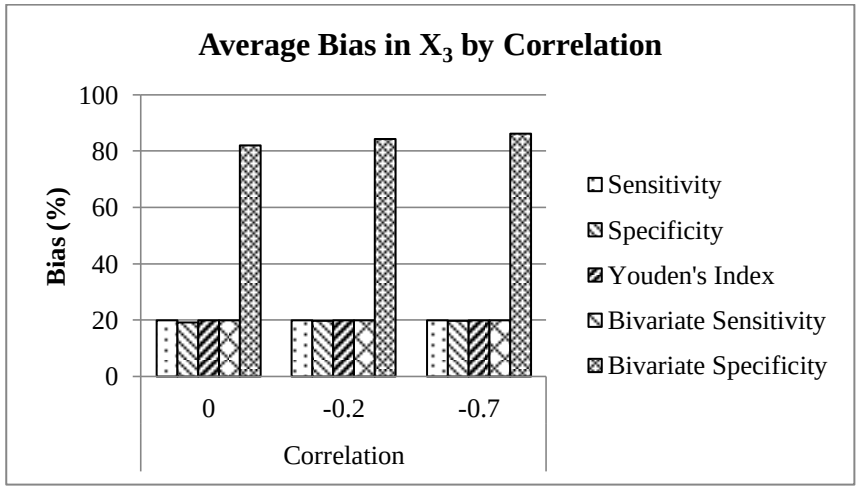
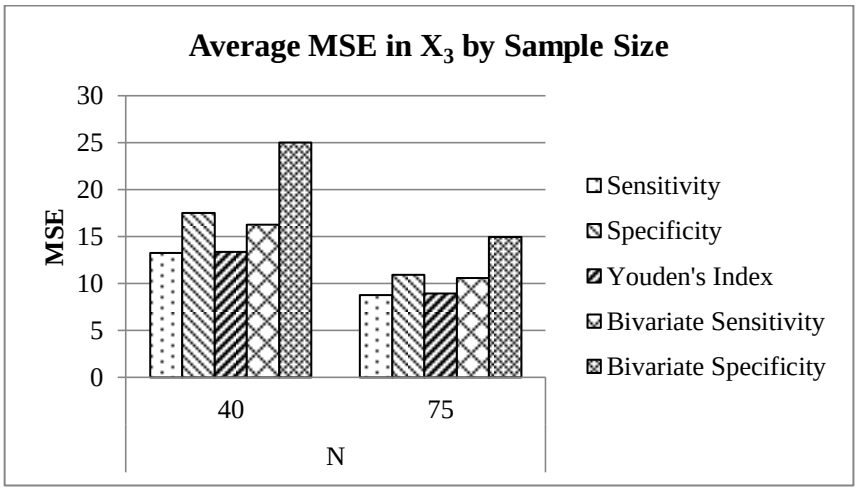
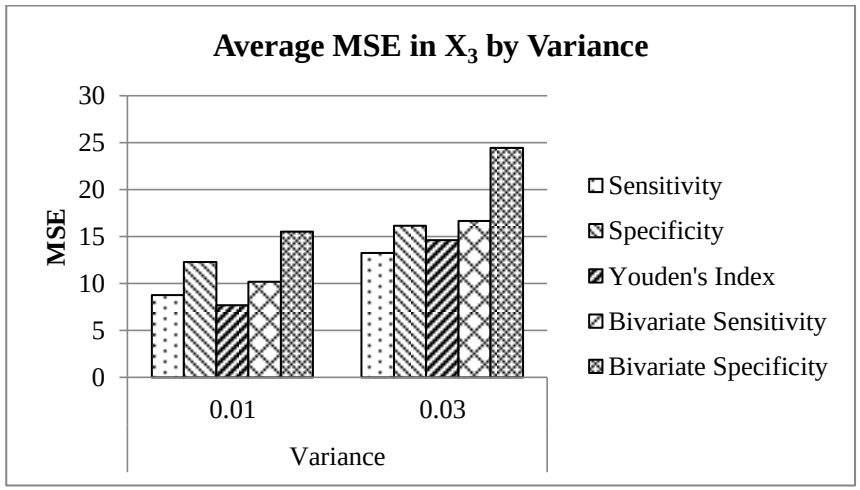


Figure B.10: Average mean square error (MSE) in X_3 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

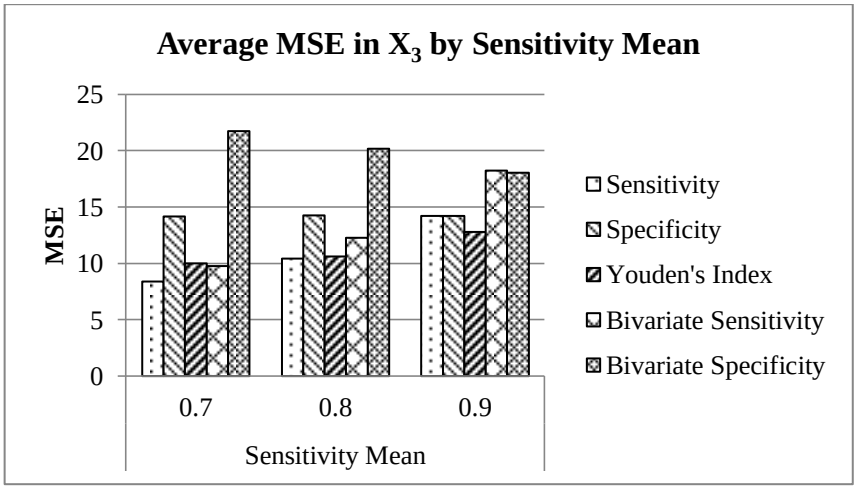
a)



b)



c)



d)

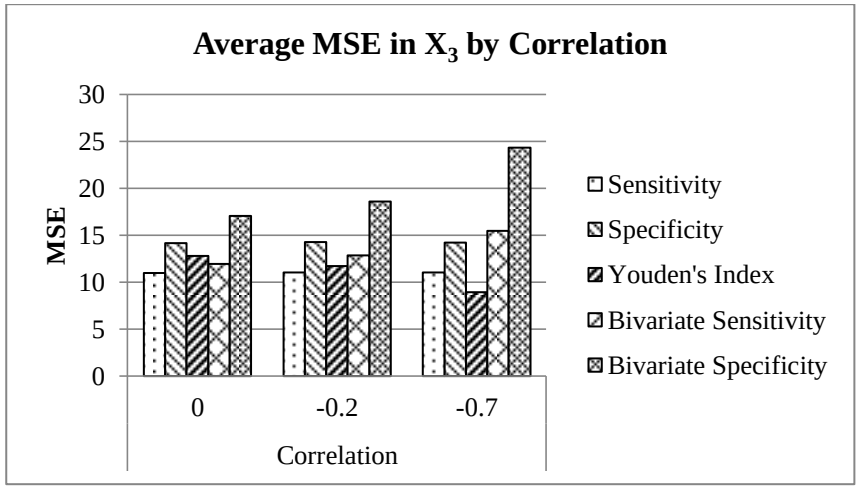
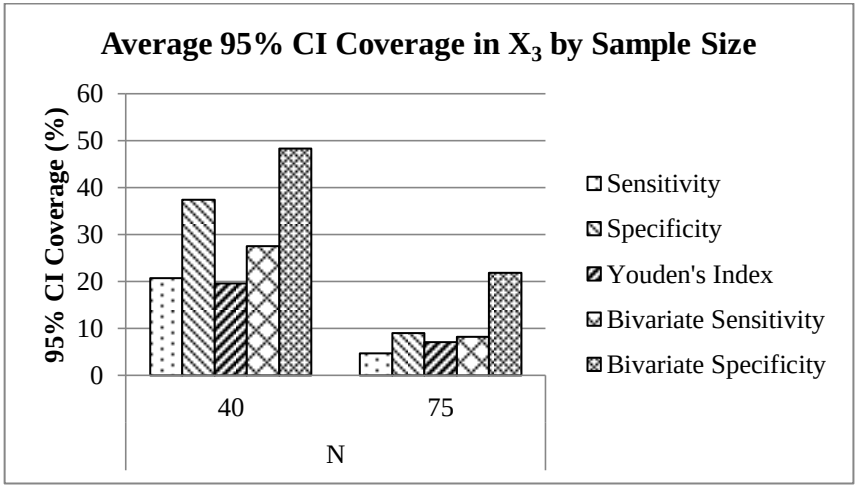
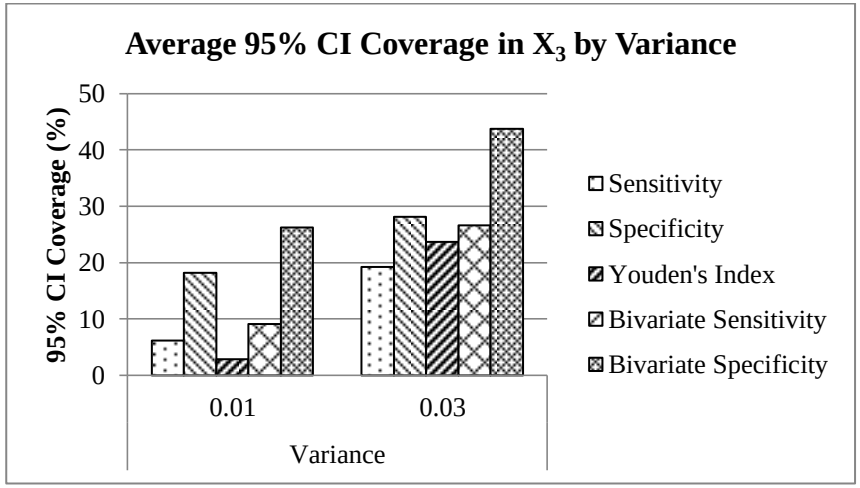


Figure B.11: Average 95% confidence interval (CI) coverage in X_3 for all models, stratified by: a) sample size, b) variance, c) sensitivity mean, and d) correlation

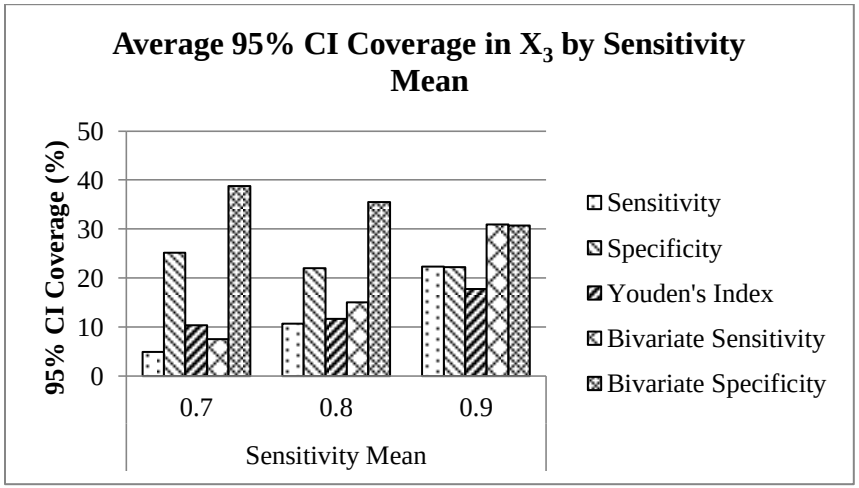
a)



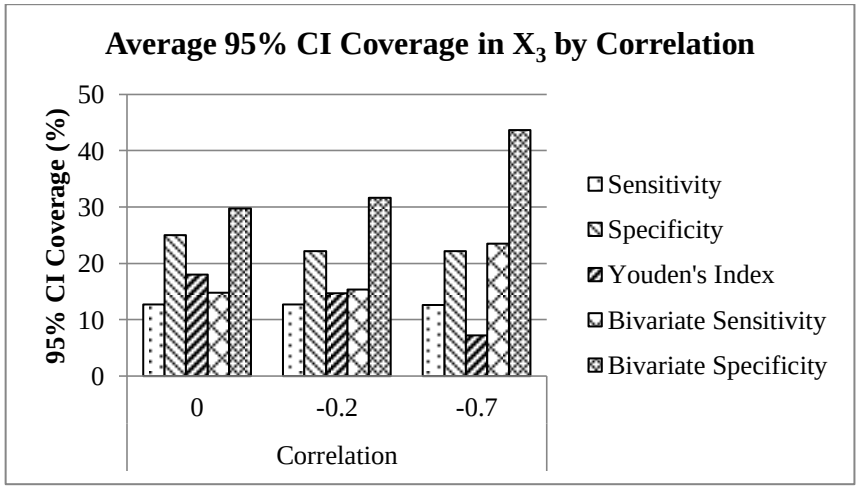
b)



c)



d)



APPENDIX C: SAS Simulation Program

A sample simulation program is provided for the univariate and bivariate models with all three binary covariates. The same program was adjusted to models with no covariates, binary covariates X1 and X2, and binary covariates X1 and X3.

```
/*************************************************************************
Modeling Diagnostic Validity Estimates from Administrative Health Data
Programmer: Kristine Kroeker
May 30, 2016
If you have further requests, you can send an email to umeinar6@myumanitoba.ca.
*************************************************************************/
%include "C:\<path>\RandMVBinary.sas"; /*importing RandMVBinary module*/
Proc iml;
/***************************/
/***Part 1 - defining modules***/
/***************************/
Load module=_all_; /*loading RandMVBinary module */
/*Module to simulate 2 correlated beta distributions (sensitivity and specificity) and 3 correlated
binary distributions (covariates X1, X2, and X3)*/
start corrbeta(N, A1, B1, A2, B2, pho); /*N=sample size, A1 and B1=shape parameters of
sensitivity, A2 and B2=shape parameters of specificity, and pho=correlation between beta
distributions*/
  /*simulate a bivariate normal distribution with specified correlation pho (0 or -.2 or -.7)*/
  Z = randNormal(N, {0, 0}, pho);
  /*transform the normal variates into uniform variates*/
  U = cdf('Normal', Z);
```

```
/*obtain beta variates for each column of U by applying the inverse beta CDF*/

Y1 = betainv(U[,1], A1, B1);

Y2 = betainv(U[,2], A2, B2);

Youden = Y1 + Y2 - 1;

/*Subtract 0.00001 when sensitivity, specificity, or Youden's index is equal to 1; then the values
  will be included in the beta distribution*/

do i=1 to N;

 if Y1[i,]=1 then Y1[i,]=Y1[i,]-.00001;

 if Y2[i,]=1 then Y2[i,]=Y2[i,]-.00001;

if Youden[i,]=1 then Youden[i,] = Youden[i,] - 0.00001;

end;

/*When Youden's index is less than zero set the value to 0.00001 for the value to be included in
  the beta distribution*/

do i=1 to N;

        if Youden[i,]<=0 then Youden[i,]=0.00001;

end;


/*Simulate 3 correlated binary distributions for covariates*/

p={0.40 0.60 0.70}; /*defining proportions for X1, X2, and X3 binary covariates*/

R={1     0.20  0.05,  /*defining correlations between X1, X2, and X3 binary covariates*/

      0.20  1     -0.2,

      0.05  -0.2  1};

X = RandMVBinary93(N, p, R);


/*Defining a count vector to represent the case definition number */
```

```
  def=j(N,1);

  def=1:N;

  def1=def`;

  /*Creates one matrix that includes case definition number and simulated values for sensitivity,
    specificity, Youden's index, and covariates for outputting from the module*/

  Matrix = def1 || Y1 || Y2 || Youden || X;

  return(Matrix);

finish;


/*Calculating the estimate, upper and lower bounds as proportions from the logit estimates in
   the univariate models*/

start uni_proportions(i, model_pe, L); /*i=case definition number, model_pe=model parameter
    estimates, L=Lth row in the model parameter estimates*/

    Q1 = j(1, 5);

    anti_logit = j(1,1);

    anti_logit_L = j(1,1);

    anti_logit_U = j(1,1);

    Q6 = j(1,8);

    Q1 = i || model_pe[L,]; /*keeping the first column for the intercept values and dropping the
     second column for the scale estimates*/

    anti_logit[1,1] = exp(Q1[1,2])/ (1 + exp(Q1[1,2])); /*Changing the logit estimate into a
     proportion*/

    anti_logit_L[1,1] = exp(Q1[1,4])/ (1 + exp(Q1[1,4])); /*lower bound as a proportion*/

    anti_logit_U[1,1] = exp(Q1[1,5])/ (1 + exp(Q1[1,5])); /*upper bound as a proportion*/

    Q6 = Q1 || anti_logit || anti_logit_stderr || anti_logit_L || anti_logit_U;

 return(Q6); /*module returns a vector of case definition number, estimate logit, standard error,
    lower bound logit, upper bound logit, estimate in proportion, lower bound in proportion, and
    upper bound in proportion*/
```

```
finish;


 /*Calculating the estimate, upper and lower bounds as proportions from the logit estimates in
the bivariate models*/

start biv_proportions(i, model_pe, L); /*i=case definition number, model_pe=model parameter
estimates, L=Lth row in the model parameter estimates*/

    /*Defining matrix sizes*/

    Q1 = j(1, 6);

    anti_logit = j(1,1);

    anti_logit_L = j(1,1);

    anti_logit_U = j(1,1);

    Q6 = j(1,9);

    Q1 = i || model_pe[L,]; /*keeping the Lth row of the parameter estimates*/

    anti_logit[1,1] = exp(Q1[1,3])/ (1 + exp(Q1[1,3])); /*Changing the logit estimate into a
     proportion*/

    anti_logit_L[1,1] = exp(Q1[1,5])/ (1 + exp(Q1[1,5])); /*lower bound as a proportion*/

    anti_logit_U[1,1] = exp(Q1[1,6])/ (1 + exp(Q1[1,6])); /*upper bound as a proportion*/

    Q6 = Q1 || anti_logit || anti_logit_stderr || anti_logit_L || anti_logit_U;

  return(Q6); /*module returns a vector of case definition number, sensitivity or specificity,
     estimate logit, standard error, lower bound logit, upper bound logit, estimate in proportion,
     lower bound in proportion, and upper bound in proportion*/

finish;


/*identifying the simulations that do not converge and setting the estimates as missing*/

start nonconverge(NumSamples, Q, mok_vect);

    Q_OK=Q || mOK_vect; /*attaching the vector (mok_vect) to identify the simulations that do
     not converge*/
```

```
    do i=1 to NumSamples;
        if mOK_vect[i,1]=0 then do; /*if the model does not converge then set all estimates to
        missing*/
            Q_OK[i,2]={.}; Q_OK[i,3]={.}; Q_OK[i,4]={.}; Q_OK[i,5]={.};
            Q_OK[i,6]={.}; Q_OK[i,7]={.}; Q_OK[i,8]={.}; Q_OK[i,9]={.};
        end;
    end;
    return(Q_OK); /*returning the matrix of all simulations with those that didn't converge with
       missing values*/
finish;

/*Module for calculating bias, MSE, and 95% CI coverage in the univariate model*/
start uni_evaluation(Q_OK, P_mean, k, NumSamples); /*Q_OK=matrix of values,
      P_mean=population mean of sens spec or Youden, k=scenario number, and
      NumSamples=number of repetitions*/
  bias = j(NumSamples, 1);
  mse = j(NumSamples, 1);
  coverage=j(NumSamples , 1, .);
  sum_coverage = j(1,1);
  /*Calculating the model evaluation statistics*/
  evaluation = j(1,6, 0); /*vector of statistics to evaluate the scenario*/
  evaluation[1,1] = mean(Q_OK[,9])*100; /*percent of simulations that converged*/
   do i=1 to NumSamples;
    if Q_OK[i,6] ^=. then  /*if the model convereged bias, mse, and coverage will be assessed*/
        bias[i,1] = abs(Q_OK[i,6]-P_mean[k,1]); /*bias is calculated from the sample mean
          estimate (proportion) minus the population mean estimate*/
        else bias[i,1] = .;
```

```
  if Q_OK[i,6] ^=. then

      mse[i,1] = ((Q_OK[i,6]-P_mean[k,1])##2) + (Q_OK[i,3]##2); /*mse is calculated from
        the sample mean estimate (proportion), population mean estimate, and the sample
        standard error*/

       else mse[i,1] = .;

       if Q_OK[i,9] ^=0 then do;

        if Q_OK[i,7] < P_mean[k,1] & P_mean[k,1] < Q_OK[i,8] then coverage[i,1]=1;
          /*Counting the number of true means that fall between the 95% CI*/

        else coverage[i,1]=0;

       end;

 end;

 sum_coverage = sum(coverage);

 evaluation[1,2] = mean(bias); /*Calculating the mean and variance of bias*/

 evaluation[1,3] = var(bias);

 evaluation[1,4] = mean(mse); /*Calculating the mean and variance of mse*/

 evaluation[1,5] = var(mse);

 evaluation[1,6] = (sum(coverage[,1])/sum(Q_OK[ ,9]))*100; /*95% confidence interval
  coverage*/

return(evaluation); /*returning the vector of proportion converged, bias mean, bias variance,
  mse mean, mse variance, and 95% CI coverage*/

finish;


/*Module for calculating bias, MSE, and 95% CI coverage in the bivariate model*/

start biv_evaluation(Q_OK, P_mean, k, NumSamples); /*Q_OK=matrix of values,
  P_mean=population mean of sens spec or Youden, k=scenario number, and
  NumSamples=number of repetitions*/

 bias = j(NumSamples, 1);

 mse = j(NumSamples, 1);
```

```
coverage=j(NumSamples , 1, .);

/*Calculating the model evaluation statistics*/

evaluation = j(1,6, 0); /*vector of statistics to evaluate the scenario*/

evaluation[1,1] = mean(Q_OK[,10])*100; /*percent of simulations that converged based on the
  binary variable indicating if the model converged*/

do i=1 to NumSamples;

 if Q_OK[i,7] ^=. then /*if the model convereged bias, mse, and coverage will be assessed*/

      bias[i,1] = abs(Q_OK[i,7]-P_mean[k,1]); /*bias is calculated from the sample mean
        estimate (proportion) minus the population mean estimate*/

      else bias[i,1] = .;

 if Q_OK[i,7] ^=. then

     mse[i,1] = ((Q_OK[i,7]-P_mean[k,1])##2) + (Q_OK[i,4]##2); /*mse is calculated from
        the sample mean estimate (proportion), population mean estimate, and the sample
        standard error*/

      else mse[i,1] = .;

      if Q_OK[i,10] ^=0 then do;

        if Q_OK[i,8] < P_mean[k,1] & P_mean[k,1] < Q_OK[i,9] then coverage[i,1]=1;
          /*Counting the number of population means that fall between the 95% CI*/

        else coverage[i,1]=0;

      end;

 end;

evaluation[1,2] = mean(bias); /*Calculating the mean and variance of bias*/

evaluation[1,3] = var(bias);

evaluation[1,4] = mean(mse); /*Calculating the mean and variance of mse*/

evaluation[1,5] = var(mse);

evaluation[1,6] = (sum(coverage[,1])/sum(Q_OK[ ,10]))*100; /*95% confidence interval
coverage*/
```

return(evaluation); */*returning the vector of proportion converged, bias mean, bias variance, mse mean, mse variance, and 95% CI coverage*/*

finish;

/**/

/*Part 2 - defining parameters and matrix sizes*/

/**/

*/*Defining simulation parameters for 36 scenarios*/*

```
sens_mean = {0.70, 0.80, 0.90, 0.70, 0.80, 0.90,
            0.70, 0.80, 0.90, 0.70, 0.80, 0.90,
            0.70, 0.80, 0.90, 0.70, 0.80, 0.90,
            0.70, 0.80, 0.90, 0.70, 0.80, 0.90,
            0.70, 0.80, 0.90, 0.70, 0.80, 0.90,
            0.70, 0.80, 0.90, 0.70, 0.80, 0.90};
spec_mean= {0.90, 0.90, 0.90, 0.90, 0.90, 0.90,
           0.90, 0.90, 0.90, 0.90, 0.90, 0.90,
           0.90, 0.90, 0.90, 0.90, 0.90, 0.90,
           0.90, 0.90, 0.90, 0.90, 0.90, 0.90,
           0.90, 0.90, 0.90, 0.90, 0.90, 0.90,
          0.90, 0.90, 0.90, 0.90, 0.90, 0.90};
Youden_mean = {0.60, 0.70, 0.80, 0.60, 0.70, 0.80,
              0.60, 0.70, 0.80, 0.60, 0.70, 0.80,
              0.60, 0.70, 0.80, 0.60, 0.70, 0.80,
              0.60, 0.70, 0.80, 0.60, 0.70, 0.80,
             0.60, 0.70, 0.80, 0.60, 0.70, 0.80,
```

```
        0.60, 0.70, 0.80, 0.60, 0.70, 0.80};
 vari = {0.01, 0.01, 0.01, 0.03, 0.03, 0.03,
      0.01, 0.01, 0.01, 0.03, 0.03, 0.03,
      0.01, 0.01, 0.01, 0.03, 0.03, 0.03,
      0.01, 0.01, 0.01, 0.03, 0.03, 0.03,
      0.01, 0.01, 0.01, 0.03, 0.03, 0.03,
      0.01, 0.01, 0.01, 0.03, 0.03, 0.03};
 N36 = {40, 40, 40, 40, 40, 40,
      40, 40, 40, 40, 40, 40,
      40, 40, 40, 40, 40, 40,
      75, 75, 75, 75, 75, 75,
      75, 75, 75, 75, 75, 75,
      75, 75, 75, 75, 75, 75};
 X1_mean = {0.40};
 X2_mean = {0.60};
 X3_mean = {0.70};
/*calculating sensitivity and specificity beta shape parameters*/
 sens_alpha = j(36,1);
 sens_gamma = j(36,1);
 spec_alpha = j(36,1);
 spec_gamma = j(36,1);
 do i=1 to 36;
   sens_alpha[i,1] = (((1-sens_mean[i,1])/vari[i,1]) - (1/sens_mean[i,1])) *
(sens_mean[i,1]*sens_mean[i,1]);
   sens_gamma[i,1] = (((1/sens_mean[i,1])-1)*sens_alpha[i,1]);
```

```
   spec_alpha[i,1] = (((1-spec_mean[i,1])/vari[i,1]) - (1/spec_mean[i,1])) *
(spec_mean[i,1]*spec_mean[i,1]);

   spec_gamma[i,1] = (((1/spec_mean[i,1])-1)*spec_alpha[i,1]);

 end;

 NumSamples = 2000; /*Number of replications*/

/*Defining size of matrices and vectors for the main program*/

 sens_Q = j(Numsamples, 8);

 sens_X1_Q = j(NumSamples, 8);

 sens_X2_Q = j(NumSamples, 8);

 sens_X3_Q = j(NumSamples, 8);

 sens_Q6 = j(1,8);

 sens_X1_Q6 = j(1,8);

 sens_X2_Q6 = j(1,8);

 sens_X3_Q6 = j(1,8);

 spec_Q = j(Numsamples, 8);

 spec_X1_Q = j(NumSamples, 8);

 spec_X2_Q = j(NumSamples, 8);

 spec_X3_Q = j(NumSamples, 8);

 spec_Q6 = j(1,8);

 spec_X1_Q6 = j(1,8);

 spec_X2_Q6 = j(1,8);

 spec_X3_Q6 = j(1,8);

 Youden_Q = j(Numsamples, 8);

 Youden_X1_Q = j(NumSamples, 8);

 Youden_X2_Q = j(NumSamples, 8);
```

```
Youden_X3_Q = j(NumSamples, 8);
Youden_Q6 = j(1,8);
Youden_X1_Q6 = j(1,8);
Youden_X2_Q6 = j(1,8);
Youden_X3_Q6 = j(1,8);
biv_sens_Q = j(Numsamples, 9);
biv_sens_X1_Q = j(NumSamples, 9);
biv_sens_X2_Q = j(NumSamples, 9);
biv_sens_X3_Q = j(NumSamples, 9);
biv_sens_Q6 = j(1,9);
biv_sens_X1_Q6 = j(1,9);
biv_sens_X2_Q6 = j(1,9);
biv_sens_X3_Q6 = j(1,9);
biv_spec_Q = j(Numsamples, 9);
biv_spec_X1_Q = j(NumSamples, 9);
biv_spec_X2_Q = j(NumSamples, 9);
biv_spec_X3_Q = j(NumSamples, 9);
biv_spec_Q6 = j(1,9);
biv_spec_X1_Q6 = j(1,9);
biv_spec_X2_Q6 = j(1,9);
biv_spec_X3_Q6 = j(1,9);
sens_mok_vect=j(NumSamples,1);
spec_mok_vect=j(NumSamples,1);
youden_mok_vect=j(NumSamples,1);
biv_mok_vect=j(NumSamples,1);
```

```
/*Defining the three correlation matrices for correlating sensitivity and specificity*/
pho1 = {1 0, 0 1};
pho2 = {1 -0.2, -0.2 1};
pho3 = {1 -0.7, -0.7 1};

/*********************************/
/*Part 3 - main simulation program*/
/*********************************/
do k=1 to 1; /*loop to go through all 36 scenarios by setting k to the scenario number*/
  Q2 = j(N36[k,1], 7);
  if k=1 | k=2 | k=3 | k=4 | k=5 | k=6 then pho=pho1;
  if k=7 | k=8 | k=9 | k=10 | k=11 | k=12 then pho=pho2;
  if k=13 | k=14 | k=15 | k=16 | k=17 | k=18 then pho=pho3;
  if k=19 | k=20 | k=21 | k=22 | k=23 | k=24 then pho=pho1;
  if k=25 | k=26 | k=27 | k=28 | k=29 | k=30 then pho=pho2;
  if k=31 | k=32 | k=33 | k=34 | k=35 | k=36 then pho=pho3;
  call randseed(14109666); /*update random seed with each scenario*/
  do i=1 to NumSamples; /*loop for each repetition*/
    Q2= corrbeta(N36[k,1], sens_alpha[k,1], sens_gamma[k,1], spec_alpha[k,1],
       spec_gamma[k,1], pho); /*create the simulated data for the scenario and randomseed*/
    /*converting the matrix to a dataset for use in SAS procedures*/
    create data from Q2[colname = {'definition' 'sens' 'spec' 'Youden' 'X1' 'X2' 'X3'}];
          append from Q2;
    close data;
    /*Fitting the univariate and bivariate models to the simulated data*/
```

```
 submit / ok=sens_mOK; /* Ok statement allows the program to continue running if the
    model does not converge*/

ods listing close;

proc glimmix data=data method=quad;

    model sens= X1 X2 X3/ dist=beta s cl ddfm=bw;

    ods output ParameterEstimates=sens_Paramest; /*saves the parameter estimates to a
       dataset called paramest*/

quit; ods output close; ods listing;

endsubmit;

submit / ok=spec_mOK; /* Ok statement allows the program to continue running if the model
    does not converge*/

ods listing close;

proc glimmix data=data method=quad;

    model spec= X1 X2 X3/ dist=beta s cl ddfm=bw;

    ods output ParameterEstimates=spec_Paramest; /*saves the parameter estimates to a
       dataset called paramest*/

quit; ods output close; ods listing;

endsubmit;

submit / ok=youden_mOK; /* Ok statement allows the program to continue running if the
    model does not converge*/

ods listing close;

proc glimmix data=data method=quad;

    model youden= X1 X2 X3/ dist=beta s cl ddfm=bw;

    ods output ParameterEstimates=youden_Paramest; /*saves the parameter estimates to a
       dataset called paramest*/

quit; ods output close; ods listing;

endsubmit;
```

```
submit/ ok=biv_mok;
ods listing close;
/*transforming the data from wide format to long format for sensitivity and specificity values
  to each represent their own row*/
Data data_long;
  set data;
  status= 1 /*sensitivity*/; diag_meas=sens; output;
  status= 2 /*specificity*/; diag_meas=spec; output;
run;
/*running the bivariate model with covariates X1, X2, and X3*/
proc glimmix data=data_long method=quad;
  class definition status;
  model diag_meas=status status*X1 status*X2 status*X3 / dist=beta noint s cl ddfm=bw;
  random status / subject=definition type=chol G;
  ods output ParameterEstimates=biv_Paramest; /*saves the parameter estimates to a dataset
    called paramest*/
quit; ods output close; ods listing;
endsubmit;
/*vectors indicating the repetitions that did not converge with a value of 0*/
sens_mok_vect[i,]=sens_mok;
spec_mok_vect[i,]=spec_mok;
youden_mok_vect[i,]=youden_mok;
biv_mok_vect[i,]=biv_mok;
/*changing the outputted parameter estimates into a matrix*/
use sens_paramest;
```

```
  read all var {effect estimate stderr lower upper};
        close sens_paramest;
    sens_model_pe= estimate || stderr || lower || upper; /*matrix has 5 rows, the 1-intercept, 2-X1
      coefficient, 3-X2 coefficient, 4-X3 coefficient, 5-scale*/
    use spec_paramest;
  read all var {effect estimate stderr lower upper};
        close spec_paramest;
    spec_model_pe= estimate || stderr || lower || upper; /*matrix has 5 rows*/
    use youden_paramest;
  read all var {effect estimate stderr lower upper};
        close youden_paramest;
    youden_model_pe= estimate || stderr || lower || upper; /*matrix has 5 rows*/
    if biv_mok=1 then do; /*Only completed if the simulation converged*/
   use biv_paramest;
  read all var {effect status estimate stderr lower upper};
        close biv_paramest;
    biv_model_pe= status || estimate || stderr || lower || upper; /*matrix has 9 rows, 1- sensitivity
      intercept, 2-specificity intercept, 3-sensitivity X1 coefficient, ..., 8-specificity X3 coefficient,
      9-scale*/
 /*calculating the proportion estimates from the logit estimates and creating a matrix of
estimates from each repetition (vector)*/
    sens_Q6 = uni_proportions(i, sens_model_pe, 1);
    sens_Q[i,] = sens_Q6;
    sens_X1_Q6 = uni_proportions(i, sens_model_pe, 2);
    sens_X1_Q[i,] = sens_X1_Q6;
    sens_X2_Q6 = uni_proportions(i, sens_model_pe, 3);
    sens_X2_Q[i,] = sens_X2_Q6;
```

```
sens_X3_Q6 = uni_proportions(i, sens_model_pe, 4);
sens_X3_Q[i,] = sens_X3_Q6;
spec_Q6 = uni_proportions(i, spec_model_pe, 1);
spec_Q[i,] = spec_Q6;
spec_X1_Q6 = uni_proportions(i, spec_model_pe, 2);
spec_X1_Q[i,] = spec_X1_Q6;
spec_X2_Q6 = uni_proportions(i, spec_model_pe, 3);
spec_X2_Q[i,] = spec_X2_Q6;
spec_X3_Q6 = uni_proportions(i, spec_model_pe, 4);
spec_X3_Q[i,] = spec_X3_Q6;
youden_Q6 = uni_proportions(i, youden_model_pe, 1);
youden_Q[i,] = youden_Q6;
youden_X1_Q6 = uni_proportions(i, youden_model_pe, 2);
youden_X1_Q[i,] = youden_X1_Q6;
youden_X2_Q6 = uni_proportions(i, youden_model_pe, 3);
youden_X2_Q[i,] = youden_X2_Q6;
youden_X3_Q6 = uni_proportions(i, youden_model_pe, 4);
youden_X3_Q[i,] = youden_X3_Q6;
biv_sens_Q6 = biv_proportions(i, biv_model_pe, 1);
biv_sens_Q[i,] = biv_sens_Q6;
biv_sens_X1_Q6 = biv_proportions(i, biv_model_pe, 3);
biv_sens_X1_Q[i,] = biv_sens_X1_Q6;
biv_sens_X2_Q6 = biv_proportions(i, biv_model_pe, 5);
biv_sens_X2_Q[i,] = biv_sens_X2_Q6;
biv_sens_X3_Q6 = biv_proportions(i, biv_model_pe, 7);
```

```
    biv_sens_X3_Q[i,] = biv_sens_X3_Q6;
    biv_spec_Q6 = biv_proportions(i, biv_model_pe, 2);
    biv_spec_Q[i,] = biv_spec_Q6;
    biv_spec_X1_Q6 = biv_proportions(i, biv_model_pe, 4);
    biv_spec_X1_Q[i,] = biv_spec_X1_Q6;
    biv_spec_X2_Q6 = biv_proportions(i, biv_model_pe, 6);
    biv_spec_X2_Q[i,] = biv_spec_X2_Q6;
    biv_spec_X3_Q6 = biv_proportions(i, biv_model_pe, 8);
    biv_spec_X3_Q[i,] = biv_spec_X3_Q6;
    end;
  end;
/*changing the estimates to missing for the repetitions that did not converge*/
  sens_Q_OK = nonconverge(NumSamples, sens_Q, sens_mok_vect);
  sens_X1_Q_OK = nonconverge(NumSamples, sens_X1_Q, sens_mok_vect);
  sens_X2_Q_OK = nonconverge(NumSamples, sens_X2_Q, sens_mok_vect);
  sens_X3_Q_OK = nonconverge(NumSamples, sens_X3_Q, sens_mok_vect);
  spec_Q_OK = nonconverge(NumSamples, spec_Q, spec_mok_vect);
  spec_X1_Q_OK = nonconverge(NumSamples, spec_X1_Q, spec_mok_vect);
  spec_X2_Q_OK = nonconverge(NumSamples, spec_X2_Q, spec_mok_vect);
  spec_X3_Q_OK = nonconverge(NumSamples, spec_X3_Q, spec_mok_vect);
  youden_Q_OK = nonconverge(NumSamples, youden_Q, youden_mok_vect);
  youden_X1_Q_OK = nonconverge(NumSamples, youden_X1_Q, youden_mok_vect);
  youden_X2_Q_OK = nonconverge(NumSamples, youden_X2_Q, youden_mok_vect);
  youden_X3_Q_OK = nonconverge(NumSamples, youden_X3_Q, youden_mok_vect);
  biv_sens_Q_OK = nonconverge(NumSamples, biv_sens_Q, biv_mok_vect);
```

```
biv_sens_X1_Q_OK = nonconverge(NumSamples, biv_sens_X1_Q, biv_mok_vect);
biv_sens_X2_Q_OK = nonconverge(NumSamples, biv_sens_X2_Q, biv_mok_vect);
biv_sens_X3_Q_OK = nonconverge(NumSamples, biv_sens_X3_Q, biv_mok_vect);
biv_spec_Q_OK = nonconverge(NumSamples, biv_spec_Q, biv_mok_vect);
biv_spec_X1_Q_OK = nonconverge(NumSamples, biv_spec_X1_Q, biv_mok_vect);
biv_spec_X2_Q_OK = nonconverge(NumSamples, biv_spec_X2_Q, biv_mok_vect);
biv_spec_X3_Q_OK = nonconverge(NumSamples, biv_spec_X3_Q, biv_mok_vect);
/*Calculating the evaluation statistics*/
sens_evaluation=uni_evaluation(sens_Q_OK, sens_mean, k, NumSamples);
sens_X1_evaluation = uni_evaluation(sens_X1_Q_OK, X1_mean, 1, NumSamples);
sens_X2_evaluation = uni_evaluation(sens_X2_Q_OK, X2_mean, 1, NumSamples);
sens_X3_evaluation = uni_evaluation(sens_X3_Q_OK, X3_mean, 1, NumSamples);
spec_evaluation=uni_evaluation(spec_Q_OK, spec_mean, k, NumSamples);
spec_X1_evaluation = uni_evaluation(spec_X1_Q_OK, X1_mean, 1, NumSamples);
spec_X2_evaluation = uni_evaluation(spec_X2_Q_OK, X2_mean, 1, NumSamples);
spec_X3_evaluation = uni_evaluation(spec_X3_Q_OK, X3_mean, 1, NumSamples);
youden_evaluation=uni_evaluation(youden_Q_OK, youden_mean, k, NumSamples);
youden_X1_evaluation = uni_evaluation(youden_X1_Q_OK, X1_mean, 1, NumSamples);
youden_X2_evaluation = uni_evaluation(youden_X2_Q_OK, X2_mean, 1, NumSamples);
youden_X3_evaluation = uni_evaluation(youden_X3_Q_OK, X3_mean, 1, NumSamples);
biv_sens_evaluation=biv_evaluation(biv_sens_Q_OK, sens_mean, k, NumSamples);
biv_sens_X1_evaluation = biv_evaluation(biv_sens_X1_Q_OK, X1_mean, 1, NumSamples);
biv_sens_X2_evaluation = biv_evaluation(biv_sens_X2_Q_OK, X2_mean, 1, NumSamples);
biv_sens_X3_evaluation = biv_evaluation(biv_sens_X3_Q_OK, X3_mean, 1, NumSamples);
biv_spec_evaluation=biv_evaluation(biv_spec_Q_OK, spec_mean, k, NumSamples);
```

```
 biv_spec_X1_evaluation = biv_evaluation(biv_spec_X1_Q_OK, X1_mean, 1, NumSamples);

 biv_spec_X2_evaluation = biv_evaluation(biv_spec_X2_Q_OK, X2_mean, 1, NumSamples);

 biv_spec_X3_evaluation = biv_evaluation(biv_spec_X3_Q_OK, X3_mean, 1, NumSamples);

/*Print the evaluation statistics*/

 print sens_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE mean'
   'MSE variance' '95% CI coverage'}];

 print sens_X1_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print sens_X2_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

print sens_X3_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print spec_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE mean'
   'MSE variance' '95% CI coverage'}];

 print spec_X1_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print spec_X2_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'};

 print spec_X3_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print youden_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print youden_X1_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print youden_X2_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print youden_X3_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];

 print biv_sens_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
   mean' 'MSE variance' '95% CI coverage'}];
```

```
  print biv_sens_X1_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
    mean' 'MSE variance' '95% CI coverage'}];

  print biv_sens_X2_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
    mean' 'MSE variance' '95% CI coverage'}];

  print biv_sens_X3_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
    mean' 'MSE variance' '95% CI coverage'}];

  print biv_spec_evaluation [colname = {'percent converged' 'bias mean' 'bias variance' 'MSE
    mean' 'MSE variance' '95% CI coverage'}];

  print biv_spec_X1_evaluation [colname = {'percent converged' 'bias mean' 'bias variance'
    'MSE mean' 'MSE variance' '95% CI coverage'}];

  print biv_spec_X2_evaluation [colname = {'percent converged' 'bias mean' 'bias variance'
    'MSE mean' 'MSE variance' '95% CI coverage'}];

  print biv_spec_X3_evaluation [colname = {'percent converged' 'bias mean' 'bias variance'
    'MSE mean' 'MSE variance' '95% CI coverage'}];

 end;

quit;
```