2015

Controlling for Confounding when Association is Quantified by Area Under the ROC Curve

Hadiza I. Galadima
Virginia Commonwealth University, issakagaladh@vcu.edu

Controlling for Confounding when Association is Quantified by Area Under the ROC Curve

A dissertation submitted in partial fulfillment of the requirement for the degree of

Doctor of Philosophy at Virginia Commonwealth University

By

Hadiza Issaka Galadima

Bachelor of Science in Statistics and Actuarial Science

St. Cloud State University, 2008, St. Cloud Minnesota

Advisor: Donna K. McClish, Ph.D.

Professor, Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia

June 2015

ACKNOWLEDGMENTS

I am deeply indebted to my advisor, Dr. Donna McClish for her mentorship and guidance. I sincerely appreciate her valuable feedback, patience, encouragement and for always being there when I need help, both personally and professionally. I would also like to thank my committee members: Dr. Christine Schubert Kabban, Dr. Wen Wan, Dr. Juan Lu, Dr. Nitai Mukhopadhyay and Dr. Roy Sabo for their support and expertise each brought to this dissertation research. Thank you to my work supervisor, Dr. April Kimmel, who never failed to ask about my dissertation work as well as my personal life, and for allowing me to take time away from work to research and write.

I am forever grateful to my late father who didn't live long enough to see my accomplishments. He taught me to be true, hard-working, and has always inspired me in pursuing higher education. I am very thankful to my mother, Hadjia Rabi Bagaya, who never fail to unconditionally support me on all the decisions I made in my life. Her constant guidance, love, prayers and strictness has helped me become who I am today. I am also thankful to my son Ben Yacine, who made me realize how beautiful life is and has been my primary motivation for the past 11 years to succeed. My siblings, Lawali, Hadjia Mariama, Elhadj Moustapha, Fati, Salamatou, Ibrahim, and Abdel Aziz who have always been there to offer a helping hand and have been most caring in all these years. I thank my step father, Ibrahim Coulibaly Seriba, who gave me valuable advice and a lot of encouragement to pursue my passion. I owe a special thanks to my best friend, Mahamadou Lawali (Bomino), for his constant support, his willingness to let me complain and for offering a shoulder to cry on when I was in low spirits. Lastly, many thanks to all my friends, aunts, uncles, cousins, in-laws, nieces and nephews whose patience and understanding, made the pursuit of my passion possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

CONTROLLING FOR CONFOUNDING WHEN ASSOCIATION IS QUANTIFIED BY AREA UNDER THE ROC CURVE

By Hadiza I. Galadima, Ph.D.

A dissertation submitted in partial fulfillment of the requirement for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2015

Advisor:
Donna K. McClish, Ph.D.
Professor, Department of Biostatistics

In the medical literature, there has been an increased interest in evaluating association between exposure and outcomes using nonrandomized observational studies. However, because assignments to exposure are not done randomly in observational studies, comparisons of outcomes between exposed and non-exposed subjects must account for the effect of confounders. Propensity score methods have been widely used to control for confounding, when estimating exposure effect. Previous studies have shown that conditioning on the propensity score results in biased estimation of odds ratio and hazard ratio. However, there is a lack of research into the performance of propensity score methods for estimating the area under the ROC curve (AUC). In this dissertation, we propose AUC as measure of effect when outcomes are continuous. The AUC is interpreted as the probability that a randomly selected non-exposed subject has a better response than a randomly selected exposed subject. The aim of this research is to examine methods to control for confounding when association between exposure and outcomes is

quantified by AUC. We look at the performance of the propensity score, including determining the optimal choice of variables for the propensity score model. Choices include covariates related to exposure group, covariates related to outcome, covariates related to both exposure and outcome, and all measured covariates. Additionally, we compare the propensity score approach to that of the conventional regression approach to adjust for AUC. We conduct a series of simulations to assess the performance of the methodology where the choice of the best estimator depends on bias, relative bias, mean squared error, and coverage of 95% confidence intervals. Furthermore, we examine the impact of model misspecification in conventional regression adjustment for AUC by incorrectly modelling the covariates in the data. These modelling errors include omitting covariates, dichotomizing continuous covariates, modelling quadratic covariates as linear, and excluding interactions terms from the model. Finally, a dataset from the shock research unit at the University of Southern California is used to illustrate the estimation of the adjusted AUC using the proposed approaches.

CHAPTER 1: INTRODUCTION

In epidemiologic research, investigators are often interested in comparing a group of people with a specific exposure to a similar group of people without that specific exposure before disease appearance or other health outcomes. This objective is easily achieved in experimental studies where the assignment to the exposure is controlled by the investigator and is done in a random fashion. More generally, the exposure group could also be treatment or non-treatment groups, populations with the risk factor or not-with the risk factor, diseased or non-diseased populations, or some other binary indicator of a clinical state. However, experimental studies are not always feasible for ethical, practical or financial reasons. For instance, in a study comparing men and women in terms of health results, gender is the exposure of interest and it is clearly impossible to randomly assign subject to different gender groups. Hence, the subjects assigned themselves to one of the exposure groups in a non-random manner; this is referred to as an observational study.

There has been an increased interest in observational studies to evaluate association between exposure (risk factors of outcome) and outcomes. Because assignments to exposure are not random in observational studies, any comparisons of outcomes between exposed and non-exposed subjects must account for factors related to the exposure of interest. This is important because failing to adjust for the confounding variables could lead to biased estimates of true effects. As a result, researchers using observational data are required to use advanced statistical methods to control for bias and confounding.

Propensity score methods have been used for a long time to reduce bias in observational studies (Rosenbaum & Rubin, 1984) . The main goal of the propensity scores is to balance

observed covariates between two groups in nonrandomized trials so that the two groups are comparable in the sense that their baseline covariates are expected to have similar distribution. The most common ways of using propensity score to reduce confounding are: stratification on the propensity score, matching on the propensity score and covariate adjustment on the propensity score (Austin, 2008; Austin, 2010; Rosenbaum & Rubin, 1983) However, a common concern in the development of propensity scores models, is the choice of variables to include in the model. So far, there is no agreed upon 'correct' propensity score model among researchers.

In his seminal work on propensity scores, Peter Austin had investigated the performance of propensity scores methods to estimate relative risk, odds ratio, hazard ratio, marginal odds ratio, marginal hazard risk and difference in means (Austin, 2007a; Austin, 2008; Austin, 2010, 2013; Austin, Grootendorst, Normand, & Anderson, 2007). However, there is no mention in the literature of the performance of propensity score when association is quantified by the area under the receiver operating characteristic (ROC) curve.

In clinical research with continuous outcomes, the area under the ROC curve (AUC) has gained an interest to assess treatment effects (Acion, Peterson, Temple, & Arndt, 2006; Brumback, Pepe, & Alonzo, 2006; Hauck, Hyslop, & Anderson, 2000). The AUC can be interpreted as the probability that a randomly selected participant in the exposed group has a larger response (or greater suspicion in terms of continuous outcome) than a randomly selected participant in the non-exposed group. For example, in a clinical study of whether or not obesity is a risk factor for hypertension, an AUC of 0.76 may imply that a randomly selected patient from the obese group (exposed group) has 76% chance of, say, a more suspicious (higher) blood pressure than a randomly selected patient from the non-obese (non-exposed) group. Here, higher values of blood pressure indicate hypertension.

This dissertation research has two major parts. In the first part, we proposed the propensity score methodology to control for confounding when association between exposure and outcomes is quantified by area under the ROC curve. Additionally, we sought to determine the optimal choice of variables to include in the propensity score model. Choices include covariates related to risk group, covariates related to outcome, covariates related to both risk group and outcome, and all measured covariates. We also compared the performance of the propensity score approach to control for confounding to that of a conventional regression approach to adjust for AUC. In the second part of this research, we examined the impact of model misspecification in AUC regression adjusting for covariates by incorrectly modelling the covariates in the data. These modelling errors include omitting covariates, dichotomizing continuous variables, modelling quadratic covariates as linear, and excluding interactions terms from the model.

This research is organized as follows. Chapter 2 contains a thorough review of literature on propensity score methods, issues of variable selection in propensity score models, the area under the ROC curve as a measure of association, methods to adjust for covariates and model misspecification issues in AUC regression analysis. Chapter 3 addresses the first part of the dissertation through a simulation study. In Chapter 4, the issue of model misspecification when estimating the AUC adjusting for covariates is investigated through a simulation study. In Chapter 5, the proposed approaches are applied to data from the Shock Research Unit at the University of Southern California, Los Angeles, California. Chapter 6 is a concluding chapter which summarizes the results of the simulations studies, addresses limitations and suggests future study. Appendices cover SAS codes to estimate the AUC controlling for confounding and validation of the simulated data along with balance diagnostics.

CHAPTER 2: BACKGROUND & REVIEW

2.1 Introduction to propensity score

2.1.1 Definition of the propensity score

In cohort studies, failing to adjust for confounding variables could lead to biased estimates of risk effect. In 1983, Rosenbaum and Rubin introduced the concept of propensity scores as a tool to reduce bias in observational studies. In randomized experiments, subjects are assigned randomly to treatment or control groups so that the two groups are comparable in the sense that the distribution of their baseline covariates are expected to be the same. However, in nonrandomized trials, the absence of random assignment doesn't guarantee a similarity in the distributions of the covariates between two groups; thus, direct comparisons may be misleading. The goal of the propensity scores, then, is to balance observed covariates between two groups in nonrandomized trials. Rosenbaum and Rubin defined the propensity score as the conditional probability of assignment to a particular group given a vector of observed covariates (Rosenbaum & Rubin, 1984). For instance, suppose each subject in the cohort has a vector of observed covariates $\boldsymbol{X}$, and an indicator of risk status $\boldsymbol{Z}$ such that $Z=1$ if subject has the risk factor and $Z=0$ if subject has no-risk factor. Then the propensity score, $e(x)=\Pr(Z=1|\boldsymbol{X})$ is the probability that a subject with covariates $\boldsymbol{X}$ is in the risk factor group.

In a randomized trial, $\Pr(Z=1|\boldsymbol{X})=\Pr(Z=0|\boldsymbol{X})$ i.e. subjects have the same chance to be assigned to treatment or control using a randomization mechanism. In this manner, the propensity score $e(x)=1/2$ for every $\boldsymbol{X}$. On the other hand, in an observational study, some subjects are more likely than others to have the risk factor or to not have the risk factor at all, so

their propensity scores could be either $e(x) > 1/2$ or $e(x) < 1/2$. This is mostly due to the non-existence of random assignment in observational studies. For example, suppose the risk factor of interest is obesity and the outcome is high blood pressure, in a study. Some subjects who are physically inactive are more likely to be in the risk group than subjects who are physically active, hence their propensity score is $e(x) > 1/2$. Another example is diabetes as a risk factor and cardiovascular disease as outcome. Subjects with no history of diabetes in the family would more likely fall in the non-risk group more often than people with a family history of diabetes; hence subjects with no family history of diabetes have a smaller probability to be in the risk group; so their propensity score would be $e(x) < 1/2$. Now, two subjects with the same propensity score, say $e(x) = 0.75$ are compared. Although, these subjects may differ in terms of their respective covariates $\boldsymbol{X}$ but the good thing is that both subjects have the same chance of being assigned to the risk group. Hence, this suggests that in the absence of random assignment, if subjects in the risk factor and non-risk factor groups are grouped or matched based on the same propensity scores, then the subjects in each group are expected to have similar covariates distributions. Therefore, the propensity score is an instrument that balances observed covariates between two risk groups in order to create the same probability structures as that achieved by a "randomized" experiment.

2.1.2 Estimating propensity scores

Several approaches exist to estimate a propensity score such as the classification tree technique using the recursive portioning and the neural networks methods (Setoguchi, Schneeweiss, Brookhart, Glynn, & Cook, 2008; Stone, Obrosky, Singer, Kapoor, & Fine, 1995) , discriminant

analysis (D'Agostino, 1998; Rosenbaum & Rubin, 1984), and the generalized additive models (Woo, Reite, & Karr, 2008). However, logistic regression is used far more often than any of the above mentioned methods. Logistic regression models the probability of having the risk factor as a function of a set of the observed covariates $\boldsymbol{X}$. The propensity score is then computed as the expected probability of being in the risk group, conditional on $\boldsymbol{X}$. The choice of covariates to be included in the propensity score model is addressed in more detail in Section 2.2.

2.1.3 Propensity score methods to estimate risk effect

Once the propensity score has been estimated, it is used as a variable in an analysis to control for confounding when estimating risk effect. The most common propensity score analysis methods include stratification, matching, and covariate adjustment on the propensity score (Austin, Grootendorst, & Anderson, 2007).

The basic idea of stratifying on the propensity score is to group subjects usually into five approximately equal-size groups determined by the quintiles of the estimated propensity score. These groups are considered to be homogeneous as subjects in each group are expected to have similar propensity scores. The use of five strata is common because researchers have shown that five groups can remove over 90% of the bias due to each baseline covariate (Cochran, 1968; Rosenbaum & Rubin, 1984). The risk effect is then estimated within each stratum. The overall estimated risk effect for the outcome will be a weighted average of the five stratum-specific risk effects. The propensity score in stratification is very useful in adjusting for baseline differences because outcome responses from the two risk groups are compared within subjects with similar propensity score. Therefore, with stratifying on the propensity score, we expect to compare individuals in risk and no-risk factor groups with similar distributions of baseline covariates $\boldsymbol{X}$.

In propensity score matching, the idea is to create matched pairs of risk factor and non-risk factor subjects with similar propensity scores. In the literature, the most commonly used matching method is the so-called greedy matching. As noted in Rosenbaum & Rubin (1984), the greedy matching includes the: a) nearest available matching on the estimated propensity score; b) Mahalanobis metric matching including the propensity score; and c) nearest available Mahalanobis metric matching within calipers defined by the propensity score. These methods have been meticulously defined elsewhere (D'Agostino, 1998; Rosenbaum & Rubin, 1985), and, therefore, they are not shown here. According to Rosenbaum the third method i.e. the greedy matching using calipers of a specified width produces the best balance between the covariates in the two risk groups (Rosenbaum, 1995). This method consists of finding a match for a randomly selected subject in the risk group by selecting the closest subject in the non-risk group within a fixed distance i.e. the predetermined caliper of the propensity score. If there are several candidates as potential match for the risk subject, then one is selected at random. If there are no candidates, for instance if no subject in the non-risk group has a propensity score close to that of the risk group subject, then the subject in the risk group is not included in the final matched sample. The process is then repeated and once the risk subject has been matched to a non-risk subject, then the latter is no longer available for consideration as a match for subsequent subjects in the risk group; this is referred to as one-to-one (1-1) matching or matching without replacement.

In the literature, users of greedy matching have matched risk groups subjects using calipers of width ranging from 0.005 to 0.01 on the propensity score scale (Austin, 2009a). However, from the results of a simulation study, Austin recommended using calipers of width 0.2 of the standard deviation of the logit of the propensity score or of width 0.02 or 0.03 on the

propensity score scale, as they tend to have superior performance compared with other methods that are used in the medical literature (Austin, 2007b, 2009a).

Other methods of matching include matching with replacement where the selected subject in the non-risk group can serve as a match for more than one subject in the risk group. However, this method has not been discussed much in the literature. Another alternative to the greedy matching is optimal matching as described by Rosenbaum in his book Observational Studies. This method consists of minimizing the total difference between the propensity scores of the risk and non-risk subjects. This method can be computationally involved and is rarely use in epidemiologic studies (Rosenbaum, 1995).

The propensity score covariate adjustment method, also referred to as regression (covariance) adjustment was described by Rosenbaum & Rubin (1983) in their early work. In this method, the outcome is regressed on two independent variables: an indicator variable $\boldsymbol{Z}$ denoting the risk status group and the estimated propensity score. The estimated risk effect is obtained from the regression coefficient for risk status. In a systematic review conducted by Weitzen et al. (2004), they have shown that over half of the selected studies used the covariate adjustment method. In these studies, the propensity score is used as either a single variable in the regression model or with additional variables in a multivariable model. In other cases, the propensity score was used as a categorical variable by dividing the propensity score into quintiles to create categories (Weitzen, Lapane, Toledano, Hume, & Mor, 2004).

Regardless of the propensity score analysis method used, the focus should be to create balance on all patients' characteristics before comparing response outcomes for patients with the risk factor and without the factor. Therefore, the estimated propensity score should be assessed

on its performance in creating balance before carrying out any outcome analysis as described in Section 2.1.4.

2.1.4 Balance diagnostics for the propensity scores

Once the strata and the matched sample based on the propensity score have been constructed, it is of great importance to check whether balance is achieved in measured baseline covariates between risk factor and non-risk factor subjects. Methods to assess balance of each covariate after propensity score adjustment include: i) measuring the standardized differences where it has been suggested that a standardized difference greater than 0.1 is considered as an important difference in the mean or prevalence of a covariate between risk factor groups (Austin & Mamdani, 2006; Normand et al., 2001); ii) assessing the distribution of the propensity scores via box plots: If the distributions of the propensity scores for risk and non-risk groups within each quintile are similar, then a good balance is achieved. Furthermore, one should assess the overall distribution of the propensity scores within each risk group (via box plots or histograms), and if they overlap then the two groups are comparable in the sense of covariates; iii) finally, one could report t-tests of equality of means between the two risk groups in regard to each continuous covariate and a chi-square test for the dichotomous covariates within each quintile to show similarity of the distribution of measured baseline covariates after propensity score adjustment. To compare baseline characteristics between exposure groups, the standardized differences have been suggested to be better than doing statistical tests as the former are independent of sample size and estimates how many standard deviations the two groups differ by (Austin, 2009c). If balance is not satisfied researchers recommend modification of the propensity score model by deleting or adding covariates or even by considering a more complex model that includes

interactions or nonlinear terms. This being said, in Section 2.2, we investigate the issues of variable selection in the propensity score models.

2.2 Issues of variables selection in propensity score models

In the literature, a common concern in developing a propensity score model is to choose which variables to include in the model. Little is known about the problem of variable selection for propensity score models (Brookhart et al., 2006). In a propensity score model, the indicator of risk status is treated as a dependent variable whereas the observed covariates are considered to be the predictors. Based on their association with the risk group and the outcome, one can categorize the observed covariates into four groups: 1) baseline covariates related to risk group; 2) baseline covariates related to the outcome; 3) baseline covariates related to both risk group and outcome; these are referred to as true confounders; 4) and finally, all measured baseline covariates (Austin, 2007a; Austin, 2008; Austin & Mamdani, 2006).

While there is no agreed upon method for determining the 'correct' propensity score model, Weitzen et al. suggest using an algorithmic method such as backward elimination, forward selection or stepwise selection for inclusion criteria (Weitzen et al., 2004). However, Monte Carlo simulations studies have shown that a propensity score model with only covariates associated with outcome or the true confounders resulted in a larger number of matched pairs, thus, resulting in a smaller bias in the estimated risk effect (Austin, 2007a) and a smaller mean squared error (Brookhart et al., 2006). These findings are consistent with the recommendations of Rubin and Thomas that a variable related to the outcome should be included in the propensity score model even if it is not statistically significant (Rubin & Thomas, 1996). The simulations studies also noted that matching on models that contain baseline covariates related to risk group

only or all measured covariates, resulted in a lower number of matched pairs (Austin, 2007b) and increased the variance of the estimated risk effect without decreasing bias (Brookhart et al., 2006). Also, D'Agostino & D'Agostino recommend "fitting a model… that includes a subset of patient characteristics that are thought to be the most important known potential confounders". The rationale behind this is to add precision to the effect estimate and adjust for any residual imbalances that may exist after the propensity score modelling (D'Agostino & D'Agostino, 2007).

It is also important to note that the findings for the choice of variables described above resulted from methods where investigators looked at outcome measures such as difference in means or proportions, odds ratios, relative risk, or hazard ratios.

Regardless of the recommendations to select the 'best' propensity score model, users of the propensity score analysis method seem to agree that the best model is based on whether balance is achieved on all baseline covariates in order to correctly estimate risk effect between patients with risk factor and non-risk factor.

2.3 Area under the ROC Curve as Measure of Association

2.3.1 P(X > Y) in clinical trials

For normal continuous outcomes, the mean difference between two populations is a well-known measure of treatment effect. However, there is an increasing interest in the literature about the use of the probability that a randomly selected participant in the treatment group (X) has a better response than a randomly selected participant in the placebo group (Y), i.e. $P(X > Y)$. The use of $P(X > Y)$ as a measure of the effect in clinical trials has been introduced by Hauck et al.

(Hauck et al., 2000) following a work by O'Brien (1988) in considering $P(X > Y)$ to assess treatment effects after noting that standard tests may fail to identify important treatment differences. Hauck et al. (2000) believe that $P(X > Y)$ is more understandable for the evaluation of treatment comparisons. They also feel that it doesn't make sense to restrict statistical approaches to the simple difference of means between two populations because these two populations might have different variations. For instance, a new treatment may have effects on the distribution of responses other than on the average response. Therefore, if there is an increased variability due to the effect of the new treatment, then the estimated effect is attenuated when $P(X > Y)$ is used as measure of treatment effect.

Acion et al. have also shown that $P(X > Y)$ is clinically more meaningful than the change in means which represents the magnitude of the mean difference but does not tell patients their chance to improve under the new treatment. They described $P(X > Y)$ as a "measure that presents good qualities of meaning, simplicity, and robustness" (Acion et al., 2006).

As noted in Tian (2008), there are a few advantages of using $P(X > Y)$ to assess treatment effects over the change in means. First, it is scale-free, making $P(X > Y)$ a reasonable measure of treatment effect no matter how much variability exists between the two populations' responses. Second, she showed that $P(X > Y)$ does not change under monotonic transformation. Hence, the theory developed for the original distribution are also valid for transformed distributions (Tian, 2008).

Furthermore, the mean difference does not account for variability within the groups being compared. Even if the standardized mean difference is used to overcome this problem, it is

difficult for clinicians to interpret practically the improvement measured in standard deviations units (Nunney, Clark, & Shepstone, 2013).

It is important to note that the probability $P(X > Y)$ is equivalent to the area under the curve (AUC) in methods for receiver operating characteristic (ROC) analysis.

2.3.2 P(X > Y) in ROC Analysis

The receiver operating characteristic (ROC) curve was originally developed for signal detection theory by Green and Swets in 1966 (Green & Swets, 1966). Since 1982, however, the ROC curves have been extensively used in the medical diagnostic testing as a powerful tool to assess how well a diagnostic test can discriminate diseased and non-diseased populations. The ROC curve is a plot of sensitivity vs 1-specificity. In general, the ROC curve describes the separation between the distribution of the continuous outcome in two different populations (Brumback et al., 2006). The ROC curve lies in the unit square, in which the diagonal line from vertices (0, 0) to (1, 1) indicates no effect i.e. the distribution of response in the disease group is the same as that of the response in the non-disease group. When the curve is pulled closer toward (0, 1) it indicates better separation of the distributions of the responses in each group. The area under the ROC curve (AUC) is an index used to summarize the accuracy of the diagnostic test. One interpretation for AUC is the probability that for a randomly selected pair of diseased and non-diseased individuals, the diagnostic test value is higher for the diseased person (Pepe, 2003).

In this dissertation research, the measure of risk effect we suggest is the probability that a randomly selected participant in the risk group has a larger response than a randomly selected participant in the non-risk group. We assume without loss of generality that larger response values are associated with the risk population, and smaller values with the non-risk population.

More generally, the risk groups could also be diseased or non-diseased populations, treatment or non-treatment (or placebo or standard treatment) groups, or some other binary indicator of a clinical state. We restrict our research to the comparison of two groups: one of subjects with the risk factor and the other of subjects without the risk factor. Let Y_{RF} and Y_{NRF} be two continuous responses from the risk and non-risk group, respectively. In ROC analysis, the area under the ROC curve (AUC) has a direct relationship with $P\left(Y_{RF} > Y_{NRF}\right)$. If there is no risk effect i.e. when the distribution of Y_{RF} is equal to the distribution of Y_{NRF}, then the AUC would be 0.50, that is $P\left(Y_{RF} > Y_{NRF}\right) = 0.5$. This probability moves toward 1 as the risk group shows a higher response.

2.4 Methods for estimating AUC

2.4.1 Correspondence of AUC with Mann-Whitney U

Let Y_{RF_i}, $(i = 1,...,n)$ and Y_{NRF_j}, $(j = 1,...,m)$ represent two continuous responses from random variables Y_{RF} and Y_{NRF} representing n subjects in the risk group and m subjects in the non-risk group, respectively. The Mann-Whitney U statistic is defined by: $U = \sum_{i=1}^{m}\sum_{j=1}^{n} I\left(Y_{RF_i} > Y_{NRF_j}\right) \Big/ mn$

where $I\left(Y_{RF_i} > Y_{NRF_j}\right)$ is an indicator function of the number of concordant pairs in which $I\left(Y_{RF_i} > Y_{NRF_j}\right) = 1 \quad if \;\; Y_{RF_i} > Y_{NRF_j}$, and 0 otherwise.

The detail of the proof of the equivalence between the Mann-Whitney statistic *U* and AUC is shown in (Pepe, 2003). However, a brief summary of this correspondence is based on the observation that $AUC = P\left(Y_{RF} > Y_{NRF}\right)$. This is evident from the ROC curve that plots

$\sum_{j=1}^{m} \frac{I\left(Y_{RF} > c\right)}{n}$ *versus* $\sum_{i=1}^{n} \frac{\left(Y_{NRF} > c\right)}{m}$ i.e. the estimation of the $P\left(Y_{RF} > c\right)$ *versus* $P\left(Y_{NRF} > c\right)$

where c represents some threshold such that a participant is classified as having the risk factor when their response is greater than c.

Sometimes, the responses Y_{RF} and Y_{NRF} are related to baselines covariates. Hence, in order to accurately compare the two outcomes, adjustment for these covariates should be made.

2.4.2 Correspondence of AUC with placement values

Delong et al. introduces the idea of placement values (DeLong, DeLong, & Clarke-Pearson, 1988). The goal of the placement values is to use the distribution of the responses in the non-risk population as the reference (or control) distribution for standardizing the responses in the risk population. For instance, suppose Y_{RF_i} and Y_{NRF_j} denote responses for a sample of n subjects in the risk group and m subjects in the non-risk group, respectively. According to the Delong Method, for a response Y_{RF_i} for a subject i in the risk group, its "placement value", called $V_i\left(Y_{RF_i}\right)$, is the fraction or percentage of the responses Y_{NRF_j} in the non-risk group that it exceeds. Hence, its placement value formula is given by:

$$V_i\left(Y_{RF_i}\right) = \frac{1}{m}\sum_{j=1}^{m} \psi\left(Y_{RF_i}, Y_{NRF_j}\right) \qquad \left(i = 1, 2, ..., n\right)$$

where $\psi\left(Y_{RF_i}, Y_{NRF_j}\right)$ is an indicator variable indicating the ordering of the responses such that $\psi\left(Y_{RF_i}, Y_{NRF_j}\right) = 1$ if $Y_{RF_i} > Y_{NRF_j}$, 0 if $Y_{RF_i} < Y_{NRF_j}$, and 0.5 if $Y_{RF_i} = Y_{NRF_j}$.

Similarly, if the distribution of the responses in the risk population is set as the reference distribution, the placement value, $V_j\left(Y_{NRF_j}\right)$, for a subject j in the non-risk group is given by:

$$V_j\left(Y_{NRF_j}\right)=\frac{1}{n}\sum_{i=1}^{n}\psi\left(Y_{RF_i},Y_{NRF_j}\right) \qquad \left(j=1,2,...,m\right)$$

Where $\psi\left(Y_{RF_i},Y_{NRF_j}\right)=1$ if $Y_{NRF_j}>Y_{RF_i}$, 0 if $Y_{NRF_j}<Y_{RF_i}$, and 0.5 if $Y_{RF_i}=Y_{NRF_j}$.

The placement value concept is a familiar way of standardizing the outcome relative to the reference population distribution. For example, if a child's weight corresponds to the 75th percentile in a healthy population then its equivalent placement value is 25% (Pepe, 2003).

Pepe et al. have also extended the Delong Method to show that the set of placement values $\left(V_i\left(Y_{RF_i}\right),V_j\left(Y_{NRF_j}\right)\right)$ can be used to plot the ROC curve (Pepe & Cai, 2004; Pepe & Longton, 2005). The area under the ROC curve (AUC) is obtained by averaging the placement values: $AUC = mean\ of\ V_i\left(Y_{RF_i}\right) = mean\ of\ V_j\left(Y_{NRF_j}\right)$.

2.5 AUC controlling for covariates

In the literature, there exist some works describing how to accommodate for covariates for AUC. Recall in the context of this research, we set $AUC=P\left(Y_{RF}>Y_{NRF}\right)$ where Y_{RF} *and* Y_{NRF} are continuous responses from a risk-group and a non-risk group, respectively.

In context for the reliability of the stress-strength system, early work introduced by Reiser et al. has examined statistical inference for $P\left(Y_1>Y_2\right)$, where Y_1 and Y_2 are independent normal variates with unknown means and variances. In their model, Reiser and Guttman considered Y_1 as the strength and Y_2 as the stress where the stress is applied to the strength of a

component. As such, $P(Y_1 > Y_2)$ measures the reliability (Reiser & Guttman, 1986). In the same context of stress-strength models, Guttman et al. (1988) estimated $P(Y_1 > Y_2)$ adjusting for covariates through linear regression models with the following assumptions: Y_1 and Y_2 are normally distributed, Y_1 and Y_2 depend (linearly) on the covariates to adjust for, and there exists an equal variance between strength (Y_1) and stress (Y_2) (Guttman, Johnson, Bhattacharyya, & Reiser, 1988).

In 2003, Faraggi extended Guttman et al.'s method to examine covariate effects on AUC, assuming a parametric distribution (Faraggi, 2003). His method is based on using regression modelling to model the covariates effects on the outcomes to obtain AUC depend on covariates.

Other work to accommodate for covariates for AUC was based on nonparametric and semiparametric theories developed by Margaret Sullivan Pepe. For continuous diagnostic tests, Pepe proposed three methods based on regression analysis techniques to control for possible effects of covariates on ROC curves (Pepe, 1998).. Her second approach which is relevant to estimating AUC while adjusting for covariates consisted of estimating AUC nonparametrically using the Wilcoxon statistic. In this approach, the AUC for each covariate with level k was estimated by $\hat{\theta}_k$. Then, the expected value of $\hat{\theta}_k$ was modelled as a linear function of the covariates X, which, at level k, are denoted by X^k such that $E\left(\hat{\theta}_k\right) = b_0 + b_1 X^k$. This method may be computationally involved and complex.

Brumback et al. developed a more general approach to accommodate for covariates for the non-parametric treatment effect, $P(Y_{RF} > Y_{NRF})$. Their method mainly consists of adjusting for a discrete covariate X. Their technique can be described as follows:

1) Each level of the discrete covariate X is considered as a stratum s such as $s = 1,...,S$, where S represents the total number of strata;

2) Within each stratum s, they compute all of the 0 or 1 indicator data such that $I\left(Y_{RF_i} > Y_{NRF_j}\right) = 1$ *if* $Y_{RF_i} > Y_{NRF_j}$, and 0 otherwise;

3) The adjusted AUC is the sum of all the indicator function of the the number of concordant pairs, i.e. $I\left(Y_{RF_i} > Y_{NRF_j}\right)$ within each strata divided by the sum of the product of the number of subjects in the risk factor group and non-risk factor group in stratum s. Hence, the adjusted estimator is given by $AUC^{adj} = \sum_{s=1}^{S}\sum_{i=1}^{n^s}\sum_{j=1}^{m^s} I\left(Y_{RF_i} > Y_{NRF_j}\right) / N$ where $N = \sum_{s=1}^{S} n^s m^s$, and n^s and m^s are the number of subjects in the risk factor and non-risk factor group in stratum s, respectively (Brumback et al., 2006). The caveat of this method is that it only accomodates a single discrete covariate.

Janes et al. proposed a covariate-adjusted measure of classification accuracy called the covariate-adjusted ROC curve, or *A*ROC for accomodating for covariates in ROC analysis. The *A*ROC is a weighted average of covariate-specific ROC curves. The deriving summary indice is the area under the covariate-adjusted ROC curve, *A*AUC which is interpreted as the probability that, for a random case and control marker observation with the same covariate value, the case observation is higher than the control. Their *A*AUC can be estimated empirically or a parametric distribution can also be assumed (Janes, Longton, & Pepe, 2009) .

In this research, we applied Janes et al.'s approach in the context of epidemiologic research to compare two risk groups while controlling for confounding and where the risk effect

is quantified by the probability that a randomly selected participant in the risk group has a larger response than a randomly selected participant in the non-risk group, $P(Y_{RF} > Y_{NRF})$.

Janes et al.'s approach in estimating the AUC controlling for confounding is based on the concept of placement values (PV). For instance, let Y_{RF} and Y_{NRF} be two continuous normal responses arising from a risk factor population and a non-risk factor population, respectively. The indicator variable T denotes the populations such that $T = 1$ if the subject has the risk factor and $T = 0$ if the subject is without the risk factor. Let $\mathbf{Z}$ denotes a vector of covariates for each subject. Let us consider the population where T = 0, as the reference or control group and use the subscript NRF (Non-Risk Factor) for index-related quantities. According to Jane et al.'s method, the covariate adjusted AUC is computed following two major steps. The first consists of estimating the cumulative distribution (CDF) for the response Y_{NRF} in the control group as a function of $\boldsymbol{Z}$ (i.e. the vector of covariates of interest requiring adjustment). This is done by specifying a linear model $Y_{NRF} = \beta_0 + \boldsymbol{Z\beta_1} + \varepsilon$ in which the error term is normally distributed and the covariates act linearly on the distribution of Y_{NRF}. Then for each subject i in the risk factor group, we compute the placement values. The placement value is the standard normal CDF of $\left(Y_{RF} - \beta_0 - \boldsymbol{Z\beta_1}\right)/\sigma$, hence $PV_{RF,Z} = \Phi\left\{\left(Y_{RF} - \beta_0 - \boldsymbol{Z\beta_1}\right)/\sigma\right\}$ where $\beta_0, \beta_1,$ and σ are the regression coefficients estimates and the standard deviation of the linear model of control observations, respectively. The second major step is to estimate the adjusted AUC which is the mean of the estimated placement values: $AUC = \sum_{i=1}^{n_{RF}} PV_{RF,Z} / n_{RF}$ where n_{RF} is the number of case observations. The algorithm for estimating the adjusted AUC under a parametric assumption is summarized in the following table:

Step	Procedure Description	Model	Assumption
1	Estimate the cumulative distribution of Y in the control group as a function of $\boldsymbol{Z}$:	$Y_{NRF} = \beta_0 + \boldsymbol{Z\beta_1} + \varepsilon$	$\varepsilon \sim N\left(0, \sigma^2\right)$
2	Calculate the placement values for each subject in the risk factor group.	$PV_{RF,Z} = \Phi\left\{\left(Y_{RF} - \beta_0 - \boldsymbol{Z\beta_1}\right)/\sigma\right\}$	β_0, β_1 *and* σ are the regression coefficients estimate, and the standard deviation, respectively of the control observations. Φ is the Standard normal CDF.
3	Estimate AUC by computing the mean of the estimated placement values.	$AUC = \sum_{i=1}^{n_{RF}} PV_{RF,Z} / n_{RF}$	n_{RF} is the number of case observations.

The standard errors for the estimated AUC are obtained by bootstrapping the data. The data is resampled separately within risk and non-risk strata. The algorithm for computing the adjusted covariate AUC has been incorporated into STATA under the *comproc* command developed by Janes et al. We developed a similar algorithm in SAS called the *%aAUC* macro to estimate the adjusted AUC-See Appendix.

2.6 Model Misspecification

Several parametric, semiparametric and nonparametric methods have been proposed in estimating AUC adjusting for covariates (Brumback et al., 2006; Faraggi, 2003; Janes et al., 2009; Pepe, 1998). However, little is known about the impact of model misspecification when estimating the AUC that accommodates for covariates. Walsh (1997) investigates the robustness of the binormal assumption by specifically investigating bias associated with the estimates of AUC if the binormal assumption was violated (Walsh, 1997). In the context of stress-strength model, Greco and Ventura in a recent work recognized that model assumptions can badly affect

the estimated AUC, and propose a robust inferential procedure to address this issue (Greco & Ventura, 2011). However, none of these methods mention model misspecification in the presence of covariates.

In 1988, Lagakos investigated the effects of misspecification in linear models for measured response variables. He examined the particular case of mismodelling or discretizing a continuous variable (Lagakos, 1988) . Furthermore, in 1990, Begg and Lagakos have considered the consequences of model misspecification when the model contains misspecified forms for both exposure and covariates (Begg & Lagakos, 1990). They found that omitting a needed variable lead to a seriously biased estimates of treatment effect.

Failing to correctly model the covariates could lead to biased estimates of treatment difference in outcome. To our knowledge, no research has been carried out to investigate the effect of covariates misspecification in estimating the adjusted AUC. In this research, we use the term “misspecification” to investigate a wide range of modelling errors and its impact on the estimated AUC. These modelling errors include omitting covariates, dichotomizing continuous variables, modelling quadratic covariates as linear, and excluding interactions terms from the model. The performance of the estimated AUC is examined based on bias, relative bias, mean squared error and coverage of 95 per cent confidence intervals.

CHAPTER 3: PROPENSITY SCORE

In this chapter, we examine the performance of propensity score methods to control for confounding when AUC is used to quantify association. We estimated several adjusted AUC using different propensity score-based methods as presented in Section 2.1.3. As a secondary objective, we sought to determine the optimal choice of variables for the propensity score model. This choice includes covariates related to risk group, covariates related to the outcome, covariates related to both risk group and outcome, and all measured variables. A simulation study was conducted to evaluate and compare the propensity score methods and models.

3.1 Design of Simulation Study

Data were simulated using a framework similar to those used by Austin et al. to examine the performance of different propensity score methods and models for estimating treatment effects (Austin, 2008; P. C. Austin et al., 2007). Data are generated according to the following steps:

Step 1: Eighteen baseline covariates were randomly generated such that nine of them were dichotomous and the other nine were continuous. Each of the 18 variables varied in their association with the risk factor group and the outcome as described in the following table:

Table 3.1 Association between baseline covariates with risk group and outcome.

	Strongly associated with risk group	Moderately associated with risk group	Not associated with risk group
Strongly associated with outcome	b_1, c_1	b_2, c_2	b_3, c_3
Moderately associated with outcome	b_4, c_4	b_5, c_5	b_6, c_6
Not associated with outcome	b_7, c_7	b_8, c_8	b_9, c_9

The 12 variables $b_1, c_1, b_2, c_2, b_4, c_4, b_5, c_5, b_7, c_7,$ and b_8, c_8 are related to the risk group, while the 12 variables $b_1, c_1, b_2, c_2, b_3, c_3, b_4, c_4, b_5, c_5,$ and b_6, c_6 are related to the outcome. The 8 variables $b_1, c_1, b_2, c_2, b_4, c_4,$ and b_5, c_5 are related to both risk group the outcome and are thus confounders. The two variables b_9, c_9 are neither associated with the risk group nor with the outcome.

The association between a given variable and risk group was measured by the odds ratio. A moderate or a strong association was assumed if the presence of the given variable in the logit model increases the odds of being in the risk group by a factor of 1.5 or 2, respectively (Austin, 2009b; Monson, 1990). A moderate or a strong association was defined as the odds of having the risk factor is increased by a factor of 1.5 or 2 for binary covariates, respectively (Austin, 2009b) and 1.5 and 1.25 for continuous covariates (Austin, 2010).

Similarly, the association between outcome and a binary variable is measured with the point-biserial correlation; the association between outcome and a continuous variable is measured with the Pearson correlation. The point biserial correlation is a measure of association between a continuous variable and a binary variable. It is a special case of the Pearson correlation. The strength of the association between a given variable and an outcome is measured

with a correlation of 0.5 and 0.3 to reflect a strong and a moderate association, respectively. Cohen in 1988 , proposed these guidelines for interpreting the magnitude of correlation coefficients (Cohen, 1988). Such correlations are not unusual in epidemiologic research. For example, in a study of association between cardiovascular death rates and municipal drinking water, Schroeder (1966) reported a correlation between death rates from arteriosclerotic heart disease and hardness of municipal waters of -0.50 ($P<0.0005$) in males and of -0.36 ($p<0.005$) in females (Schroeder, 1966). Another study from the child and adolescent trial for cardiovascular health (Osganian et al., 1999) also showed correlations of magnitude similar to those considered here. For example, the study has found a strong correlation between folic acid and vitamin B_6 ($r =0.48$; $P =0.001$) and "somewhat stronger" correlation between serum homocysteine and folic acid ($r= -0.36$; $P=0.001$).

Hence, for this simulation study, we considered correlations values of 0.5, 0.3 and 0 to depict strong, moderate and no association, respectively between a given variable and the outcome; and odd ratios values of 2, 1.5, and 1 for a strong, moderate, and no association between a covariate and the risk factor group.

To determine the optimal choice of variables for the propensity score (PS) model, four propensity score models were specified in the Monte Carlo simulation experiments:

PS-Model 1: This model includes all 12 variables associated with the risk factor group:

$b_1, c_1, b_2, c_2, b_4, c_4, b_5, c_5, b_7, c_7,$ and b_8, c_8 .

PS-Model 2: This model includes all 12 variables associated with the outcome:

$b_1, c_1, b_2, c_2, b_3, c_3, b_4, c_4, b_5, c_5,$ and b_6, c_6 .

PS-Model 3: This model includes all 8 variables associated with both the risk factor group and the outcome: $b_1, c_1, b_2, c_2, b_4, c_4,$ and b_5, c_5.

PS-Model 4: This model includes all 18 generated variables: $b_1 - b_9$ and $c_1 - c_9$.

Step 2: Next, we generated a risk factor status T for each subject. To do so, data were simulated such that the logit of the probability of having the risk factor for the i[th] subject is linearly related to the 12 covariates associated with the risk factor group. In other words, the subject-specific probability of group assignment was determined assuming that the probability of group assignment $\left(P_{group}\right)$ was related to the 12 baseline covariates that are strongly and moderately associated with the risk group i.e. $\left(b_1, b_2, b_4, b_5, b_7, b_8, c_1, c_2, c_4, c_5, c_7, c_8\right)$ through the following logit model:

$$logit = \log\left(\frac{P_{group}}{1-P_{group}}\right) = \beta_0 + \beta_1 b_1 + \beta_2 b_2 + \beta_4 b_4 + \beta_5 b_5 + \beta_7 b_7 + \beta_8 b_8 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_4 c_4 + \alpha_5 c_5 + \alpha_7 c_7 + \alpha_8 c_8 \qquad (3.1)$$

Hence, the subject-specific probability of group assignment is obtained by inversing the logit:

$$P_{group} = \frac{\exp\left(logit\right)}{1+\exp\left(logit\right)} \qquad (3.2)$$

The risk factor status T for each of the N subjects was generated from a Bernoulli distribution with a parameter $\left(P_{group}\right)$ i.e. $T \sim Bernoulli\left(P_{group}\right)$. The risk factor status vector is computed by comparing the estimated probability of group assignment $\left(P_{group}\right)$ to a random variable U

generated from $Uniform(0,1)$. We assign $T = \begin{cases} 1 & \text{if } U \le P_{group} \\ 0 & \text{if } U > P_{group} \end{cases}$. T =1 if the subject has the risk factor and T=0 otherwise.

Step 3: In this last step of our data generating process, for each of the N subjects, a continuous outcome Y conditional on risk factor status T was generated through the following linear model:

$$Y = \alpha_0 + \delta T + \beta_1^* b_1 + \beta_2^* b_2 + \beta_3^* b_3 + \beta_4^* b_4 + \beta_5^* b_5 + \beta_6^* b_6 + \alpha_1^* c_1 + \alpha_2^* c_2 + \alpha_3^* c_3 + \alpha_4^* c_4 + \alpha_5^* c_5 + \alpha_6^* c_6 + \varepsilon \qquad (3.3)$$

Each regression coefficient was estimated assuming the outcome Y and the single covariate X (i.e. $b_1 - b_9, c_1 - c_9$) were related through a regression equation:

$$Y = \beta X + \varepsilon \qquad (3.4)$$

where β is a regression parameter and ε represents modelling error such that $\varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$ The covariate X could be continuous, $X \sim N(\mu_x, \sigma_X^2)$, or dichotomous, $X \sim Bernoulli(p)$. The following is a derivation of the formula used to estimate the regression coefficients in Equation (3.3)

If X and Y are linearly related, then the Pearson product-moment correlation is estimated by

$$\rho = \frac{Cov(Y, X)}{\sqrt{Var(Y) \times Var(X)}} \qquad (3.5)$$

The formula for ρ is known to be related to the regression coefficient as:

$$\rho = \beta \frac{\sigma_x}{\sigma_y} \qquad (3.6)$$

From equation *(3.6)*, we derived the regression coefficient β such that

$$\beta = \rho \frac{\sigma_y}{\sigma_x} \qquad (3.7)$$

The formula for β can be written in terms of variances of X and ε,

$$\beta = \rho \frac{\sqrt{\beta^2 \sigma_x^2 + \sigma_\varepsilon^2}}{\sigma_x} \qquad (3.8)$$

where $\sigma_y^2 = Var[Y] = Var[\beta X + \varepsilon] = \beta^2 Var[X] + Var[\varepsilon] = \beta^2 \sigma_x^2 + \sigma_\varepsilon^2$

Solving *(3.8)* for β, each regression coefficient in equation *(3.3)* is determined by

$$\beta = \rho \frac{\sigma_\varepsilon}{\sigma_x} \sqrt{\frac{1}{(1 - \rho^2)}} \qquad (3.9)$$

where σ_x and σ_ε are the standard deviations of the covariate of interest and the error term, respectively.

The effect on outcome of risk group compared to non-risk group is quantified by AUC statistic through δT in Equation *(3.3)*. Hence, the effect size is given by $\delta = \sigma_\varepsilon * \sqrt{2} \Phi^{-1}(AUC_0)$ that is δ is a function of the true AUC which is denoted AUC_0. This formula can be derived as follows:

When outcomes are normally distributed in the risk factor (RF) and non-risk (NRF) populations i.e. $Y_{RF} \sim N\left(\mu_{RF}, \sigma_{RF}^2\right)$, $Y_{NRF} \sim N\left(\mu_{NRF}, \sigma_{NRF}^2\right)$, and $\varepsilon \sim N(\mu_\varepsilon, \sigma_\varepsilon^2)$, then the AUC for the binormal ROC curve is:

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right) \qquad (3.10)$$

where $a = \frac{\mu_{RF} - \mu_{NRF}}{\sigma_{RF}}$, $b = \frac{\sigma_{NRF}}{\sigma_{RF}}$, Φ denotes the standard normal cumulative distribution function (Pepe, 2003).

We assumed $\mu_{NRF} = 0$, $\sigma_{RF} = \sigma_{NRF} = \sigma_\varepsilon$. Hence, $a = \frac{\mu_{RF}}{\sigma_\varepsilon}$ and $b = 1$.

Thus, the true AUC can be expressed as

$$AUC_0 = \Phi\left(\frac{\mu_{RF}}{\sigma_\varepsilon\sqrt{2}}\right) \qquad (3.11)$$

Solving Equation *(3-11)* for μ_{RF} which we called δ for simplicity, the effect size is given by

$$\delta = \mu_{RF} = \sigma_\varepsilon * \sqrt{2}\Phi^{-1}\left(AUC_0\right) \qquad (3.12)$$

3.2 Simulating Data

A sample of size N = 500 was considered in this simulation study; for each of the N subjects, we randomly generated:

1) 18 independent baseline covariates such that 9 of them are dichotomous variables from a

Bernouilli distribution with parameter 0.5: $(b_1, b_2, b_4, b_5, b_7, b_8, b_9) \sim Bernouilli(0.5)$ and the other 9 are continuous from a standard normal distribution: $(c_1, c_2, c_4, c_5, c_7, c_8, c_9) \sim N(0,1)$. Each of the 18 covariates varies in their association with the risk group and the outcome as described in Table 3.1.

2) A risk factor status for each of the N subjects by first generating the logit model in Equation (3.1):

$$logit = \log\left(\frac{P_{group}}{1-P_{group}}\right) = -1.65 + \log(2)b_1 + \log(1.5)b_2 + \log(2)b_4 + \log(1.5)b_5 + \log(2)b_7 + \log(1.5)b_8 + \log(1.5)c_1 + \log(1.25)c_2 + \log(1.5)c_4 + \log(1.25)c_5 + \log(1.5)c_7 + \log(1.25)c_8$$

β_0 is set to -1.65, so that approximately 50% of subjects would be exposed to the risk factor group. This was determined in an initial set of Monte Carlo simulations. As described in section 3.1, we set $(\beta_1, \beta_4, \beta_7) = \log(2)$ *and* $(\alpha_1, \alpha_4, \alpha_7) = \log(1.5)$ to depict a strong association between the risk group with the binary and continuous covariates, respectively; $(\beta_2, \beta_5, \beta_8) = \log(1.5)$ and $(\alpha_2, \alpha_5, \alpha_8) = \log(1.25)$ to depict a moderate association between the risk group with the binary and continuous variables, respectively. Next, we generated a risk factor status T according to the methods described in Section 3.1 and Equation (3.2).

3) A continuous outcome conditional on the risk factor status T using Equation (3.3):

$$Y = \delta T + 4.6b_1 + 4.6b_2 + 4.6b_3 + 2.6b_4 + 2.6b_5 + 2.6b_6 + 2.3c_1 + 2.3c_2 + 2.3c_3 + 1.3c_4 + 1.3c_5 + 1.3c_6 + \varepsilon$$

where $\varepsilon \sim N(0,4)$. The regression coefficients were determined using Equation (3.9) such that the correlation value between $(b_1, b_2, b_3, c_1, c_2, c_3)$ and the outcome would be 0.5 for a strong association and the correlation between $(b_4, b_5, b_6, c_4, c_5, c_6)$ and the outcome would be 0.3 for a moderate association. We set $\alpha_0 = 0$. δ is a function of the true AUC as shown in Equation (3.12). We considered three different values of AUC_0 in the outcomes-generating process: 0.5, 0.7, and 0.9. These values were set according to the general rule of interpreting AUC suggested by Hosmer and Lemeshow (Hosmer & Lemeshow, 2000). An AUC of 0.5 indicates no association between outcome and exposure; an AUC of 0.7 indicates an acceptable association; and an AUC of 0.9 indicates an excellent association between exposure and outcome. The AUC values of 0.5, 0.7, and 0.9 are interpreted as follows: If we randomly select two subjects, one with the risk factor and the other without the risk factor, the probability that the subject with the risk factor has the condition is 0.5, 0.7, and 0.9 respectively. In other words, there is a 50-50, 70-30 and 90-10 chance for a subject having the risk factor to develop the condition compared to a subject not having the risk factor.

The data generating process described here was repeated 2500 times. All data generation and analyses were completed using SAS version 9.3 and 9.4.

3.3 Estimating the propensity score

In this research, the propensity score is estimated using a logistic regression model where the risk factor status T is regressed on measured baseline covariates. To determine the optimal choice of variables for the propensity score model, we consider four categories of variables for inclusion in the propensity score model: 1) variables related to risk group; 2) variables related to

the outcome; 3) variables related to both risk group and outcome; and 4) all measured variables. The logistic model depicting the conditional probability of assignment to a particular risk group given a vector of observed covariates $\boldsymbol{X}_i$ for the i^{th} subject is given by:

$$\Pr\left(T_i = 1 \mid \boldsymbol{X}_i\right) = e\left(x_i\right) = \frac{e^{x_i \beta_i}}{1 + e^{x_i \beta_i}} \qquad (3.13)$$

Where T_i is the binary group assignment and $T_i = 1$ if the subject belong to the risk factor group and $T_i = 0$ if the subject is in the non-risk factor group. $\boldsymbol{\beta}_i$ is the vector of regression parameters.

3.4 Constructing strata and matched sets with the estimated propensity scores

In general, the estimates of propensity scores are used for sub-classification and in matching to control for confounding in observational studies.

As described in Section 2.1.3, a step-by-step approach described by D'Agostino Jr. (1998) and Perkins (2000) was used to create propensity score strata based on the quintiles of the estimated propensity scores (D'Agostino, 1998; Perkins, Tu, Underhill, Zhou, & Murray, 2000).

Furthermore, matched pairs of risk factor and no- risk factor subjects with similar propensity scores were formed. The 1:1 greedy matching technique using calipers of width 0.2 of the standard deviation of the logit of the propensity score was used to form these pairs (P. C. Austin et al., 2007; Austin & Mamdani, 2006). The %GMATCH macro in SAS obtained from the Mayo Clinic website at http://www.mayo.edu/research/departments-divisions/department-health-sciences-research/division-biomedical-statistics-informatics/software/locally-written-sas-macros was used to construct a SAS dataset containing the matched subjects.

3.5 Estimation of risk effect

Risk effects were estimated via the AUC statistic using different propensity score analysis methods as described in Chapter 2. We also used different propensity score models in estimating AUC to explore how variable selection affects the estimate of the risk effect on the outcome.

3.5.1 Unadjusted AUC

As described in Section 2.4.1 the unadjusted AUC is computed based on the fact that it is equivalent to the two-sample Mann-Whitney *U* statistic (Brumback et al., 2006; Mann & Whitney, 1947; Pepe, 2003) in the form:

$$U = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} I\left(Y_{RF_i} > Y_{NRF_j}\right)}{mn} \qquad (3.14)$$

$$where\ I\left(Y_{RF_i} > Y_{NRF_j}\right) = \begin{cases} 1, & if\ \ Y_{RF_i} > Y_{NRF_j} \\ 0, & otherwise \end{cases}$$

Y_{RF_i}, $(i = 1,...,n)$ and Y_{NRF_j}, $(j = 1,...,m)$ are two continuous responses from random variables Y_{RF} and Y_{NRF} representing populations in the risk group and the non-risk group, respectively. The variance of the unadjusted AUC is calculated based on a formula suggested by (DeLong et al., 1988) which is incorporated into SAS via PROC LOGISTIC.

3.5.2 AUC based on stratifying on the propensity score (The PS stratified AUC)

The adjusted AUC based on stratifying on the propensity score was obtained by extending the method proposed by Brumback et al. (2006) as described in Section 2.5. Following their technique, the adjusted AUC is given by:

$$AUC^{adj} = \sum_{s=1}^{S} \sum_{i=1}^{n^s} \sum_{j=1}^{m^s} I\left(Y_{RF_i} > Y_{NRF_j}\right) / N \qquad (3.15)$$

where $N = \sum_{s=1}^{S} n^s m^s$, and n^s and m^s are the number of subjects in the risk factor and non-risk factor group in stratum s, respectively.

Our proposed method is based on Equation (3.15) where each strata is determined by the quintiles of the estimated propensity scores; that is the 20th, 40th, 60th, 80th, and 100th percentile of the total sample. It can readily be shown that the proposed adjusted AUC that we refered to as the "Adjusted Propensity Score Stratified AUC" is a weighted average of the stratum-specific AUCs, given by:

$$AUC^{adj}_{Stratified} = \sum_{s=1}^{S} w_s AUC_S \qquad (3.16)$$

Where $w_s = \frac{m^s n^s}{\sum_{s=1}^{S} m^s n^s}$, m^s and n^s are are the number of subjects in the risk factor and non-risk factor group in stratum s, respectively. $S = (1, 2, 3, 4, 5)$ correspond to the quintiles of the propensity score.

This can be readily seen as follows. Suppose Y_{RF} and Y_{NRF} are two continuous outcome measures for m subjects in the risk factor (RF) group and n subjects in the non-risk (NRF)

group. Let X be a discrete covariate corresponding to 5 stratas where each strata is determined by the quintiles of the estimated propensity scores. From Equation (3.14), the unadjusted AUC is given by:

$$AUC^{unadj} = P\left(Y_{RF} > Y_{NRF}\right) = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I\left(Y_{RF} > Y_{NRF}\right)$$

Hence, the stratum-specific AUC can be written as:

$$AUC_S = \frac{1}{m^s n^s}\sum_{i=1}^{m^s}\sum_{j=1}^{n^s} I\left(Y_{RF} > Y_{NRF}\right) \qquad (3.17)$$

where $(s = 1, 2, 3, 4, 5)$ and m^s and n^s represent the number of subjects in the risk factor and non-risk factor group in stratum s, respectively.

From Equation (3.15), the adjusted AUC can be written as:

$$AUC^{adj}_{Stratified} = \sum_{s=1}^{S} AUC_S * \frac{m^s n^s}{\sum_{s=1}^{S} m^s n^s} \qquad (3.18)$$

We observe that the adjusted AUC in Equation (3.18) is a weighted average of the stratum-specific AUC_S. Thus, Equation (3.18) may be rewritten to obtain the Equation in (3.16):

$$AUC^{adj}_{Stratified} = \sum_{s=1}^{S} w_s AUC_S \text{ where } w_s = \frac{m^s n^s}{\sum_{s=1}^{S} m^s n^s}$$

The variance of the adjusted propensity score stratified AUC is given by:

$$Var\left[AUC^{adj}_{Stratified}\right] = \sum_{s=1}^{S}\left(w_s\right)^2 Var\left[AUC_S\right] \qquad (3.19)$$

The variance equation is derived as follows.

$$VAR\left[AUC_{Stratified}^{adj} \right] = VAR\left[\sum_{s=1}^{S} w_s AUC_S \right]$$
$$= \sum_{s=1}^{S} VAR\left[w_s AUC_S \right]$$
$$= \sum_{s=1}^{S} \left(w_s \right)^2 VAR\left[AUC_S \right]$$

The AUCs standard errors are also obtained using Delong's approach (1988) of estimating the variance of the Mann-Whitney statistic which was incorporated in SAS PROC LOGISTIC (DeLong et al., 1988).

3.5.3 AUC based on matching on the propensity score

We estimated the adjusted risk effect via AUC in the propensity score matched sample based on Janes *et al.'s* method (2009) for accommodating covariates in ROC analysis as described in Section 2.5. The estimated risk group effect is estimated in the matched sample as the mean of the placement values (PVs) for each subject with propensity score PS in the risk group:

$$AUC_{matched}^{adj} = \sum_{i=1}^{n_{RF}} PV_{RF,PS} / n_{RF} \qquad (3.20)$$

where n_{RF} is the number of subjects having the risk factor in the matched sample. The PVs of the response Y_{RF} for each subject with estimated propensity score PS in the risk group is given by: $PV_{RF,PS} = \Phi\left\{ \left(Y_{RF} - \beta_0 - \beta_1 PS \right) / \sigma \right\}$. β_0, β_1 *and* σ are the estimates of regression coefficients and the root mean squared error, respectively, from the observations in the non-risk group. These estimates were obtained through a regression model of the response Y_{NRF} in the non-risk group as a function of the propensity score PS. The model is given by $Y_{NRF} = \beta_0 + \beta_1 PS + \varepsilon$, where

$\varepsilon \sim N\left(0, \sigma^2\right)$. The variance estimates of the adjusted AUC were obtained via bootstrapping using 1000 bootstrap samples of the original observations.

3.5.4 AUC based on covariate adjustment using the propensity score

The adjusted risk group effect is estimated under the covariate adjustment method by regressing the outcome variable on the estimated propensity score and the variable representing risk group status T using the regression method developed by Janes *et al* (2009) described in Section 2.4. The standard errors for the estimated AUC were obtained by bootstrapping the data.

3.5.5 AUC based on simple regression adjustment not using the propensity score

For comparison purposes we estimated AUC based on ROC regression method (Janes et al., 2009). This method consists in directly modelling covariates effects on the response, within the general context of regression. Hence, the outcome is modelled as a function of an indicator variable denoting the risk group status and a set of independent covariates. We refer to this method as the "direct AUC regression adjustment" method. We use the same four groupings of covariates for inclusion in the regression model as were used in the propensity models: 1) covariates related to risk group; 2) covariates related to the outcome; 3) covariates related to both risk group and outcome; and 4) all measured covariates. The effect of the covariates on the outcome is directly estimated by the AUC statistic using the concept of placement values as described in Section 2.4.2.

3.6 Evaluation criteria for estimated AUC

As evaluation criteria for the performance of the estimated AUC, we considered bias, relative bias, variance estimation, mean squared error (MSE), and coverage of 95% confidence interval across the 2,500 simulated data sets. In this section we review some of those criteria.

3.6.1 Bias and relative bias

An estimation of bias of the estimated adjusted AUC for a given propensity score model is the mean estimated adjusted AUC of the 2,500 samples minus the true AUC that is used in the data generating process. The relative bias provides a measure of the magnitude of the bias; it is defined as the ratio of the estimator bias and its true value.

$$Bias = \overline{AUC} - AUC_{true} \qquad (3.21)$$

$$Relative\ Bias = 100 * \left(\frac{\overline{AUC} - AUC_{true}}{AUC_{true}} \right) \qquad (3.22)$$

3.6.2 Mean squared error and root mean squared error

The mean squared error (MSE) of the estimator AUC is the average squared difference between the estimator AUC and the true value of the risk effect AUC_{true}. It incorporates both bias and variance. The RMSE is the square root of the MSE. The more accurate estimator would lead to a smaller MSE and RMSE.

$$MSE = \frac{\sum_{i=1}^{2,500} \left(AUC - AUC_{true} \right)^2}{2,500} = Var\left(AUC \right) + \left(Bias \right)^2 \qquad (3.23)$$

3.6.3 Coverage probability of 95% confidence interval

For each propensity score method and model, a coverage probability of the 95% CI is reported. The coverage probability is the percentage of estimated 95% confidence intervals that contain the true AUC. The intervals estimators were computed using the normal approximation interval i.e. $95\%CI = \left[\overline{AUC} - t_{0.975,n-1}\, s/\sqrt{n}, \overline{AUC} + t_{0.975,n-1}\, s/\sqrt{n} \right]$ where the $(1-\alpha/2)$ quantile of the t distribution is $t_{1-\alpha/2,n-1}$ with $n-1$ degrees of freedom. Coverage was estimated by counting the proportion of times out of 2500 that the estimated confidence interval included the true value. We determined whether the coverage was significantly different from 0.95 by approximating the distribution of the binomial probability p using the normal distribution with standard deviation $\sqrt{\frac{1}{n} p(1-p)}$. Hence, the normal approximation interval is given by $\hat{p} \pm z\sqrt{\frac{1}{n} \hat{p}(1-\hat{p})}$. Therefore, based on the 2,500 simulated data, any coverage outside of $[94.15\%,\ 95.85\%]$ is statistically different from 95%.

3.7 Results of the simulation study

Results of the simulation study are given in Tables 3.2 to 3.8 and in Figures 3.1 to 3.3. The mean estimated risk effect across the 2,500 simulated data sets for each propensity sore methods and each model we considered is given in Table 3.2. The crude estimate is biased positively when the true risk group effect are 0.5 and 0.7 but is biased negatively when the true AUC is 0.9.

When stratification on the quintiles of the propensity score is used, we observe three things: 1) the amount of bias is similar within each effect group regardless of the propensity score model used; 2) the risk effect is overestimated when there is no effect (True AUC = 0.5) and underestimated when the true effects are 0.7 and 0.9; 3) the risk estimate when truth is 0.7 is associated with the least bias.

When matching on the propensity score is used, we noticed that when there is no effect, the bias is almost null, but it is not the case when the true was 07 or 0.9. Also, the choice of models did not matter, the risk estimates were similar across all models for each true AUC.

When covariate adjustment on the propensity score is used, the findings are similar to those previous ones. When AUC is 0.5, the results are similar to those found with matching, including the RMSE. However, PS model 2 i.e. model including variables associated with outcome seems to have the least bias. These findings are not consistent across the true effects and the amount of bias is still high.

From these results, it appears that stratifying, matching and covariate adjustment on the propensity score resulted in biased estimation of AUC. When true effects were 0.7 and 0.9, the estimated risks from all methods and models were negatively biased with relative biased ranging from -15% to -7% as seen in Table 3.4. Prior research demonstrated that conditioning on the propensity score resulted in biased estimation of odds ratio and hazard ratio (P. C. Austin et al., 2007). So our results are not totally unexpected.

Finally, we investigated risks effects estimated from simple regression adjustment for comparative purposes. The mean estimated risk effects perform better than those estimated from the propensity score models. The second regression model including all covariates associated with outcomes was found to be the best model in estimating the true effect. Similarly, the fourth

model including all measured covariates resulted in unbiased estimates of the risk effect except when true was 0.5. However, the first and third models which do not include all the variables related to outcome resulted in biased estimates of the true AUC. Also, these models increased MSE especially when true effects were 0.7 and 0.9.

In conclusion, if an investigator is interested in estimating AUC while controlling for covariates, we recommend not to use the propensity score methods to adjust covariates; instead the conventional AUC regression adjustment is the method to use. Furthermore, AUC regression modeling adjusting for covariates related to the outcome and the model adjusting for all variables lead to unbiased estimation of AUC.

Table 3.2 AUC Estimates from different methods and different models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	0.6293	0.7302	0.8468
PS Stratify -M1	0.5502	0.6374	0.7900
PS Stratify -M2	0.5488	0.6493	0.8078
PS Stratify -M3	0.5511	0.6511	0.8101
PS Stratify -M4	0.5495	0.6370	0.7890
PS Matching -M1	0.4967	0.6137	0.7638
PS Matching -M2	0.5042	0.6314	0.7897
PS Matching -M3	0.4901	0.6227	0.7882
PS Matching -M4	0.4986	0.6160	0.7663
PS Covariate Adjust - M1	0.5067	0.6374	0.7900
PS Covariate Adjust - M2	0.4970	0.6493	0.8078
PS Covariate Adjust - M3	0.4823	0.6511	0.8101
PS Covariate Adjust - M4	0.5117	0.6370	0.7890
Reg. Adjustment - M1	0.5070	0.6530	0.8257
Reg. Adjustment - M2	**0.5013**	**0.7018**	**0.9014**
Reg. Adjustment - M3	0.4862	0.6341	0.8128
Reg. Adjustment - M4	**0.5166**	**0.7139**	**0.9065**

Figure 3.1 AUC Estimates from different methods and different models

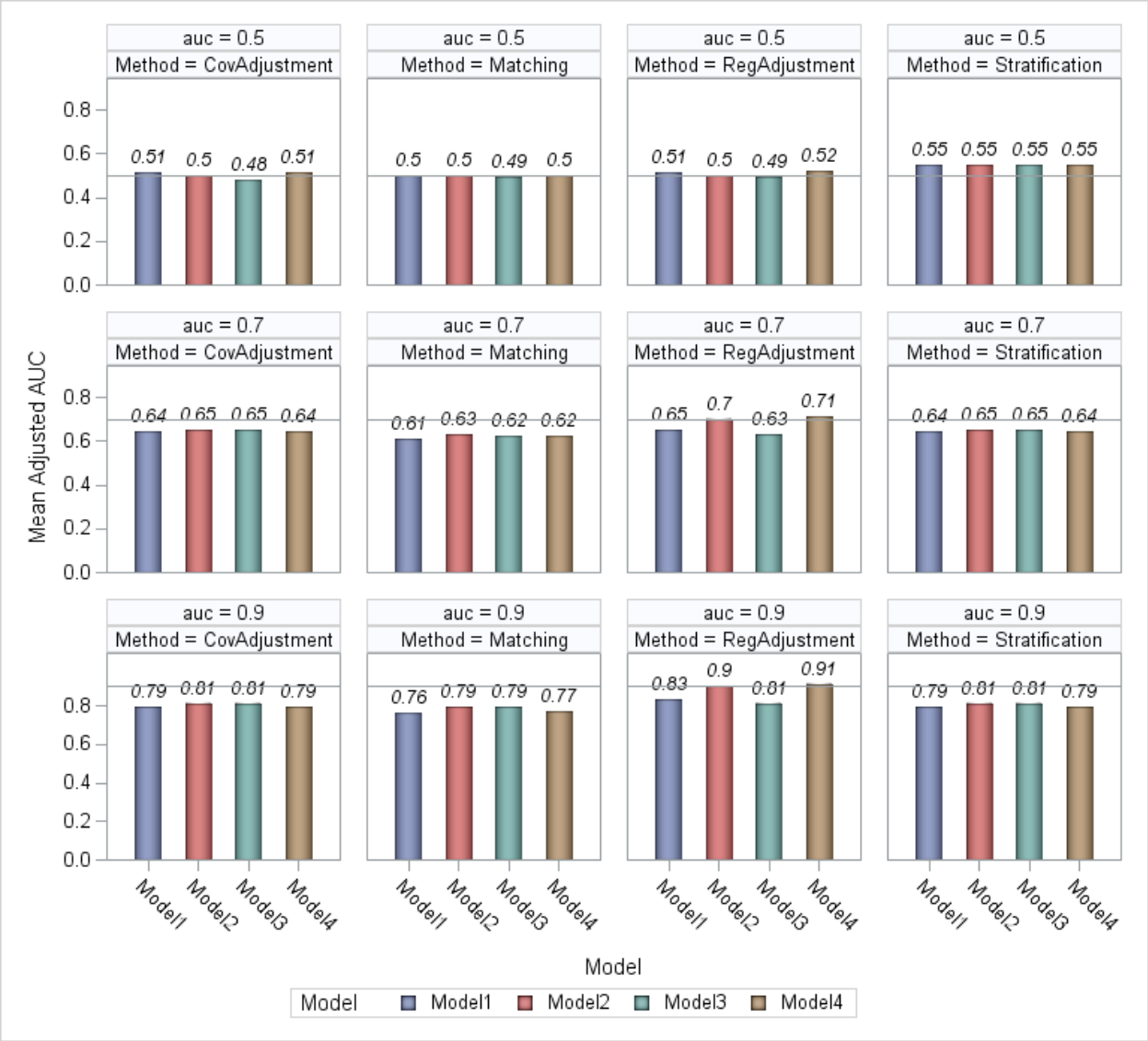


Table 3.3 Bias in estimating AUC using Different PS models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	0.1293	0.0302	0.0532
PS Stratify -M1	0.0502	-0.0626	-0.1100
PS Stratify -M2	0.0488	-0.0507	-0.0922
PS Stratify -M3	0.0511	-0.0489	-0.0899
PS Stratify -M4	0.0495	-0.0630	-0.1110
PS Matching -M1	-0.0033	-0.0863	-0.1362
PS Matching -M2	0.0042	-0.0686	-0.1103
PS Matching -M3	-0.0099	-0.0773	-0.1119
PS Matching -M4	-0.0014	-0.0840	-0.1337
PS Covariate Adjust - M1	0.0067	-0.0745	-0.1243
PS Covariate Adjust - M2	-0.0030	-0.0751	-0.1150
PS Covariate Adjust - M3	-0.0177	-0.0854	-0.1187
PS Covariate Adjust - M4	0.0117	-0.0709	-0.1227
Reg. Adjustment - M1	0.0070	-0.0470	-0.0743
Reg. Adjustment - M2	**0.0013**	**0.0018**	**0.0014**
Reg. Adjustment - M3	-0.0138	-0.0659	-0.0872
Reg. Adjustment - M4	**0.0166**	**0.0139**	**0.0065**

Table 3.4 Relative Bias in estimating AUC using different methods and different models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	25.8527	4.3076	5.9070
PS Stratify -M1	10.0387	-8.9477	-12.2259
PS Stratify -M2	9.7690	-7.2440	-10.2455
PS Stratify -M3	10.2277	-6.9815	-9.9870
PS Stratify -M4	9.9047	-8.9956	-12.3315
PS Matching -M1	-0.6688	-12.3302	-15.1365
PS Matching -M2	0.8322	-9.7960	-12.2543
PS Matching -M3	-1.9724	-11.0410	-12.4274
PS Matching -M4	-0.2886	-11.9942	-14.8570
PS Covariate Adjust - M1	1.3354	-10.6492	-13.8160
PS Covariate Adjust - M2	-0.6033	-10.7243	-12.7760
PS Covariate Adjust - M3	-3.5442	-12.1923	-13.1864
PS Covariate Adjust - M4	2.3374	-10.1260	-13.6343
Reg. Adjustment - M1	1.3956	-6.7087	-8.2597
Reg. Adjustment - M2	**0.2647**	**0.2619**	**0.1545**
Reg. Adjustment - M3	-2.7593	-9.4143	-9.6898
Reg. Adjustment - M4	3.3258	1.9813	0.7268

Figure 3.2 Relative Bias in estimating AUC using different methods and different models

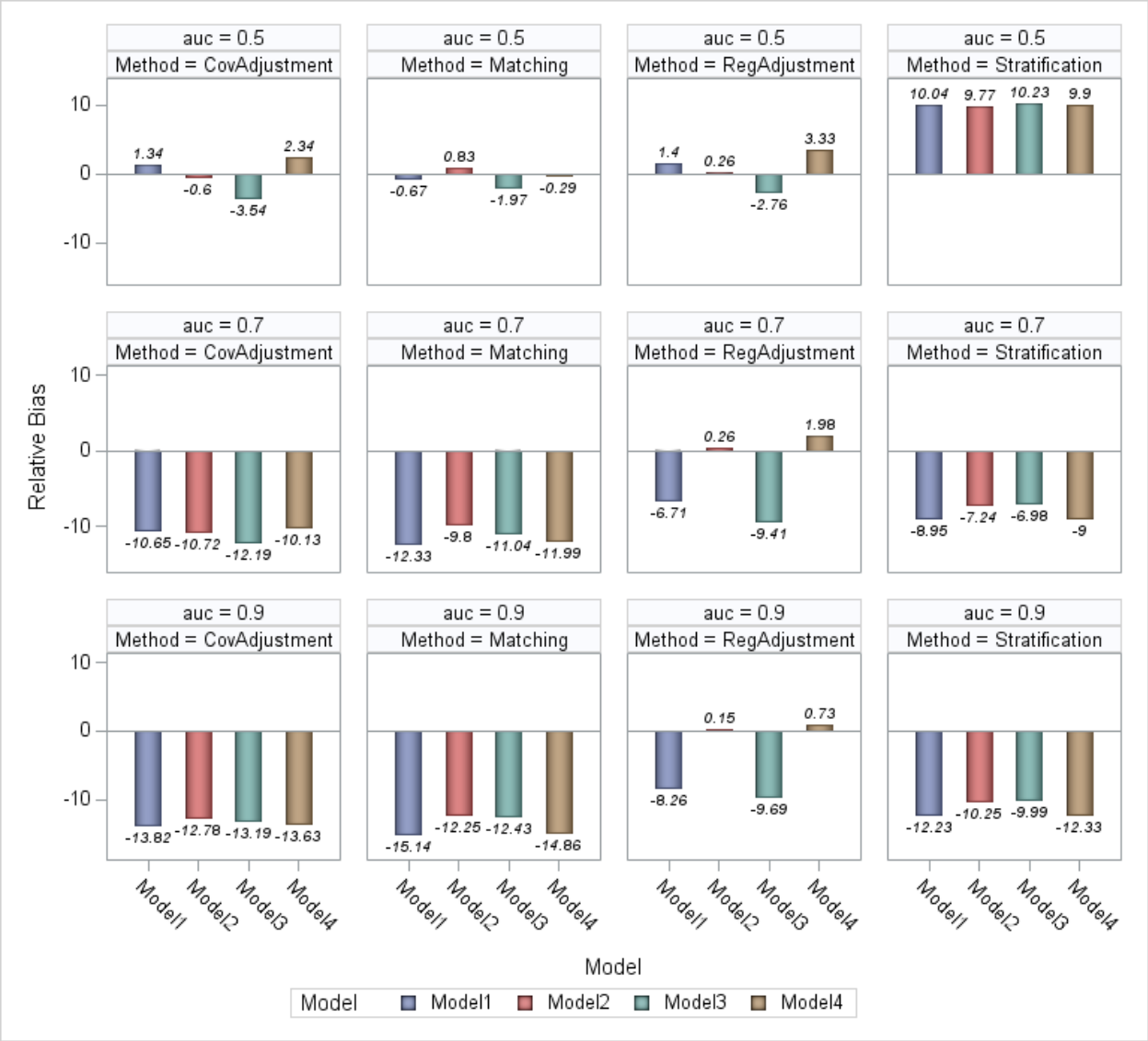


Table 3.5 Standard error in estimating AUC using Different PS models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	0.0248	0.0222	0.0170
PS Stratify -M1	0.0641	0.0609	0.0490
PS Stratify -M2	0.0614	0.0580	0.0452
PS Stratify -M3	0.0610	0.0575	0.0447
PS Stratify -M4	0.0648	0.0616	0.0497
PS Matching -M1	0.0216	0.0209	0.0174
PS Matching -M2	0.0209	0.0199	0.0159
PS Matching -M3	0.0213	0.0205	0.0164
PS Matching -M4	0.0221	0.0213	0.0177
PS Covariate Adjust - M1	0.0174	0.0167	0.0137
PS Covariate Adjust - M2	0.0173	0.0166	0.0133
PS Covariate Adjust - M3	0.0177	0.0171	0.0138
PS Covariate Adjust - M4	0.0173	0.0166	0.0136
Reg. Adjustment - M1	0.0186	0.0174	0.0130
Reg. Adjustment - M2	0.0186	0.0166	0.0096
Reg. Adjustment - M3	0.0184	0.0176	0.0133
Reg. Adjustment - M4	0.0187	0.0165	0.0094

Table 3.6 Root Mean Squared Error in estimating AUC using different methods and models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	0.1316	0.0375	0.0558
PS Stratify -M1	0.0814	0.0874	0.1205
PS Stratify -M2	0.0785	0.0770	0.1027
PS Stratify -M3	0.0796	0.0755	0.1004
PS Stratify -M4	0.0816	0.0881	0.1216
PS Matching -M1	0.0219	0.0888	0.1373
PS Matching -M2	0.0213	0.0714	0.1114
PS Matching -M3	0.0235	0.0800	0.1130
PS Matching -M4	0.0222	0.0866	0.1349
PS Covariate Adjust - M1	0.0186	0.0764	0.1251
PS Covariate Adjust - M2	0.0176	0.0769	0.1158
PS Covariate Adjust - M3	0.0251	0.0871	0.1195
PS Covariate Adjust - M4	0.0209	0.0728	0.1235
Reg. Adjustment - M1	0.0198	0.0501	0.0755
Reg. Adjustment - M2	0.0186	0.0167	0.0097
Reg. Adjustment - M3	0.0230	0.0682	0.0882
Reg. Adjustment - M4	0.0251	0.0216	0.0114

Figure 3.3 Root Mean Squared Error in estimating AUC using different methods and models

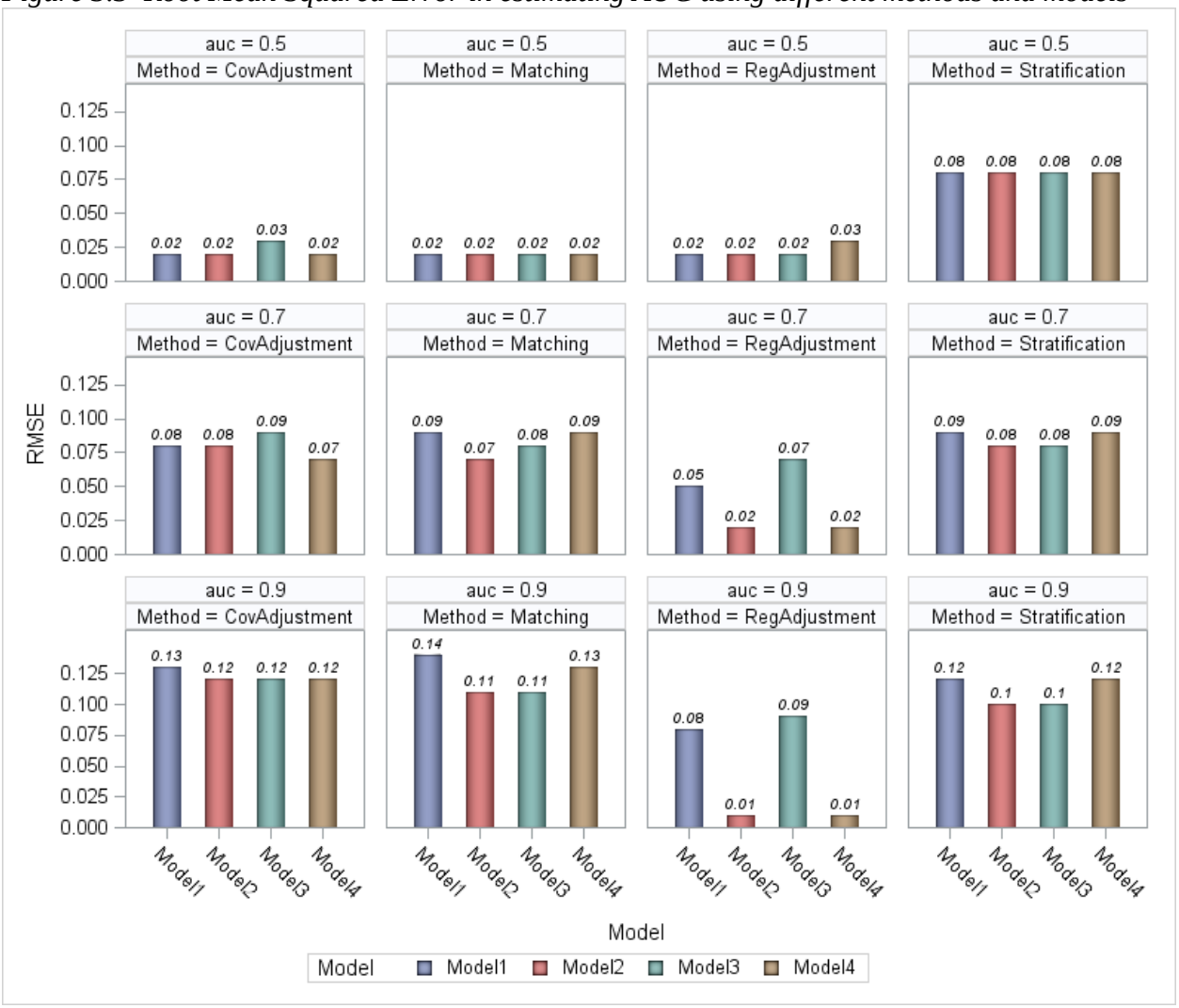


Table 3.7 Coverage of 95% confidence intervals for AUCs using different PS models

	True AUC		
Models/Methods	***0.5***	***0.7***	***0.9***
Unadjusted	0.00	71.04	9.64
PS Stratify -M1	99.96	99.60	20.92
PS Stratify -M2	99.80	99.96	41.04
PS Stratify -M3	99.76	99.64	45.40
PS Stratify -M4	99.92	99.96	18.84
PS Matching -M1	91.76	2.08	0.00
PS Matching -M2	83.56	13.60	0.00
PS Matching -M3	78.96	11.28	0.00
PS Matching -M4	95.08	2.04	0.00
PS Covariate Adjust - M1	86.24	2.36	0.00
PS Covariate Adjust - M2	75.04	6.68	0.00
PS Covariate Adjust - M3	64.88	4.80	0.00
PS Covariate Adjust - M4	83.76	3.36	0.00
Reg. Adjustment - M1	93.36	22.32	0.00
Reg. Adjustment - M2	**95.28**	**94.92**	**94.20**
Reg. Adjustment - M3	87.96	3.32	0.00
Reg. Adjustment - M4	85.60	85.76	86.76

CHAPTER 4: MODEL MISSPECIFICATION

In Chapter 3, we have shown that propensity score methods (stratification, matching and covariate adjustment) resulted in biased estimation of the true AUC. We have also shown that the direct AUC regression adjustment on the covariates lead to unbiased estimation of AUC under certain circumstances. This is true especially when the covariates related to the outcome or all measured covariates are included in the model.

The AUC regression adjustment is based on modelling the ROC curve as a function of placement values to estimate the adjusted AUC, as described in Section 2.5. However, in observational studies, little is known about the impact of misspecifying the model adjusting for the AUC. Therefore, in this part of the dissertation research, we aim to assess model misspecification in AUC regression adjustment. In other words, we sought to determine the impact of incorrectly modelling the covariate effects on the risk effect estimate. We conducted a simulation study to evaluate model misspecification in AUC regression adjustment. The simulation study is designed to specifically illustrate the following aims: 1) The impact of missing influential variables; 2) The impact of modelling continuous variables as dichotomous; 3) The impact of failing to include interactions; 4) and the impact of non-linearity.

4.1 Design of the simulation study

The data generating process is similar to those used in Chapter 3, Section 3.1. The following have been simulated:

1) Three independent continuous covariates x_1, x_2, x_3. Each of the three covariate varies in their association with the outcome and the risk factor group as described in the following table:

Table 4.1 Association between baseline covariates with risk group and outcome

	Strongly associated with risk group	Moderately associated with risk group	Weakly associated with risk group
Strongly associated with outcome	x_1		
Moderately associated with outcome		x_2	
Weakly associated with outcome			x_3

Hence, x_1 is strongly associated with the risk group and the outcome; x_2 is moderately associated with the risk group and the outcome; and x_3 is weakly associated with the risk group and the outcome. For aims 1- 3 of this simulation, the strength of the association between a given variable and the outcome is measured with the Pearson correlation. We consider correlations values of 0.7, 0.4 and 0.1 to depict strong, moderate and weak association, respectively (Pett, 1997). Similarly, the association between the risk factor group and the covariates is measured by the odds ratio. A strong, moderate or a weak association is defined as the odds of having the risk factor is increased by a factor of 4.5, 2.5, and 1.5, respectively (Rosenthal, 1996).

Table 4.2 A guide to strength of association

Association	Correlation
Strong	.7 - .89
Moderate	.5 - .69
Weak	0 - .28

Source: Adapted from Pett, 1997

2) A risk factor status T_i was generated such that the logit of the probability of having the risk factor for the i[th] subject is linearly related to (x_1, x_2, x_3). The logit model is given by:

$$logit = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (4.1)$$

where $\beta_1 = \log(4.5), \beta_2 = \log(2.5), \beta_3 = \log(1.5)$. The value of β_0 was set according to the specific aim under investigation. Hence, for each subject, a treatment status denoted by T was generated from a Bernoulli distribution with parameter (P_{treat}) where $P_{treat} = \exp\left(\frac{logit}{1+logit}\right)$, i.e. $T \sim Bernoulli(P_{treat})$. T =1 if the subject has the risk factor and T=0 otherwise. The treatment status vector is computed by comparing the estimated probability of assignment to a random variable U generated from $Uniform(0,1)$. We assigned $T = 1$ if $U \le P_{treat}$ and $T = 0$ if $U > P_{treat}$.

3) A continuous response Y_i was randomly generated as an outcome conditional on risk factor status (T_i) and a set of independent covariates. The outcome is modelled specific to the aim under investigation. More details are given in section below.

4.2 Data Simulation

A sample of size N = 500 was considered for this simulation with 1,500 replications. The data generating process was done according to the methods in section 4.1 to specifically illustrate the following four aims:

4.2.1 Aim1: The impact of missing influential variables

To examine the impact of missing influential variables, for each of the N subjects, we generated three independent covariates $(x_1, x_2, x_3) \sim Normal(0,1)$ according to Table 4-1. We generated a

treatment status for each subject based on the logit model in Equation (4.1). β_0 was set to 0 so that approximately 50% of subjects would be exposed to the risk factor group. This was determined in an initial set of Monte Carlo simulations. The true outcome model is generated using the following linear model:

$$Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon \qquad (4.2)$$

where $\varepsilon \sim N(0,4)$. The regression coefficients were determined according to Equation (3.9). Thus, $\alpha_1 = 3.92,\ \alpha_2 = 1.75,\ \alpha_3 = 0.4$. We set $\alpha_0 = 0$. The effect on outcome of risk group compared to non-risk group is quantified by the AUC statistic through δ in Equation 4-2 using the relationship $\delta = \sigma_\varepsilon * \sqrt{2}\Phi^{-1}(AUC_0)$ as described in Equation (3.12). We considered three different values of AUC_0 in the outcomes-generating process: 0.5, 0.7, and 0.9. These values were set according to the general rule of interpreting AUC suggested by Hosmer and Lemeshow (2000) to indicate no discrimination, an acceptable discrimination, and an excellent discrimination, respectively (Hosmer & Lemeshow, 2000).

Under Aim1, we specifically sought to determine what would happen if the model is missing:

a) A covariate strongly associated with the outcome. So, the investigative model is given by: $Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon_i$.

b) A covariate moderately associated with the outcome? ($Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_3 x_3 + \varepsilon_i$)

c) A covariate weakly associated with the outcome? ($Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon_i$)

d) Covariates moderately and weakly related to the outcome? ($Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \varepsilon_i$)

4.2.2 Aim2: The impact of modelling continuous covariates as dichotomous.

To evaluate the effect of modeling continuous variables as dichotomous, the data generating process is the same as in section 4.2.1 and the true outcome model is exactly similar to Equation (4.2). We investigated the following models:

a) A covariate strongly associated with the outcome is dichotomized i.e. x_1 is dichotomized as D_1 .The research model is $Y_i = \alpha_0 + \delta T + \alpha_1 D_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon_i$

b) A covariate moderately associated with the outcome is dichotomized: x_2 is dichotomized as D_2 ($Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_2 D_2 + \alpha_3 x_3 + \varepsilon_i$

c) A covariate weakly associated with the outcome is dichotomized: x_3 is dichotomized as D_3 ($Y_i = \alpha_0 + \delta T + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 D_3 + \varepsilon_i$)

d) All three covariates are dichotomized i.e. x_1, x_2, x_3 are dichotomized as D_1, D_2, D_3 $\left(Y_i = \alpha_0 + \delta T + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_3 + \varepsilon_i\right)$.

4.2.3 Aim3: The impact of excluding interactions

To evaluate the impact of missing interactions in model misspecification, we consider the following outcome model with interaction effects:

$$Y_i = \delta T_i + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_1 x_1 x_2 + \alpha_2 x_1 x_3 + \alpha_3 x_2 x_3 + \varepsilon_i \qquad (4.3)$$

where $\varepsilon \sim N\left(0,4\right)$ and $x_1, x_2, x_3 \sim Normal\left(1,1\right)$ are independent variables associated with the outcome strongly, moderately, and weakly, respectively. In this setting $x_1 - x_3$ have mean 1 rather than 0 so that they would not be centered. Not centering the variables would effectively

eliminate the association between main terms and interactions terms. We refer to the independent terms, x_1, x_2, x_3 as "main terms" and x_1x_2, x_1x_3, x_2x_3 as 'interaction terms'. To keep things simple, the coefficients of the interaction terms in Equation 4-3 were set to be the same coefficients as we would have used for the main effects for strong, moderate and low association. The regression coefficients $\alpha_1, \alpha_2, \alpha_3$ were determined using Equation 3-9. We set $\alpha_1 = 3.92$, $\alpha_2 = 1.75$, $\alpha_3 = 0.4$ and $\alpha_0 = 0$. β_0 in Equation 4.1 was set to -2.83 so that approximately 50% of the subjects would be exposed to the risk factor group. The interaction terms are likely to be correlated with the main effect terms as seen in Table 4.3; this is referred to as multicollinearity. However, we believe multicollinearity can safely be ignored in this situation as discussed by Woolridge in Introductory Econometrics (Wooldridge, 2000). Woolridge argues that collineratity induced by two main effects and their interaction (for example x1, x2 and x1x2) are not something to worry about as they are not linearly related. For instance both variables should be included in the regression to capture the relation between the predictor and the outcome as a function of another predictor.

Table 4.3 Correlation coefficients between outcome, main effects and interactions terms

Pearson Correlation Coefficients

	Y	x1	x2	x3
x1	0.7361	1	-0.0003	0.0002
x2	0.4653	-0.0003	1	0.0001
x3	0.1949	0.0002	0.0001	1
x1x2	0.8674	0.5772	0.5771	0.0001
x1x3	0.6143	0.5768	-0.0002	0.5766
x2x3	0.3985	-0.0001	0.5770	0.5770

We evaluated the impact of interaction according to these investigative models:

a) Ignoring any interaction term with the strong covariate i.e.

$$Y_i = \delta T_i + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_3 x_2 x_3 + \varepsilon_i$$

b) Ignoring any interaction term with the moderate covariate

$$Y_i = \delta T_i + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_2 x_1 x_3 + \varepsilon_i$$

c) Ignoring any interaction term with the weak covariate i.e.

$$Y_i = \delta T_i + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \alpha_1 x_1 x_2 + \varepsilon_i$$

d) Ignoring all interaction terms ($Y_i = \delta T_i + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon_i$)

4.2.4 Aim 4: The impact of non-linearity

To evaluate the performance of model misspecification on nonlinear fitting, we consider the following quadratic regression model:

$$Y_i = \delta T_i + \alpha_1 \left(x_1 - c_1\right)^2 + \alpha_2 \left(x_2 - c_2\right)^2 + \alpha_3 \left(x_3 - c_3\right)^2 + \varepsilon_i \qquad (4.4)$$

where we generated x_1, x_2, x_3 uniformly distributed from 0 to 4 and $\varepsilon_i \sim N\left(0, 2.5\right)$. The regression coefficients $\alpha_1, \alpha_2, \alpha_3$ were set to 6, 3, and 2 respectively to depict strong, moderate and weak association between the outcome and the covariates. We used the Hoeffding's D to determine the correlations between x_1, x_2, x_3 and Y. Unlike the Pearson and the Spearman correlations, the Hoeffding's D can be used to detect nonlinear dependency beyond linear and monotonic association (Hoeffding, 1948) . The values of the statistic vary between -0.5 and 1, with 1 indicating complete dependence. We also used a visual check to measure the strength of

the association between Y and the covariates. (c_1, c_2, c_3) in Equation 4-4 represent the lowest points where the quadratic curve changes direction i.e. the vertex. We consider two sets of values of (c_1, c_2, c_3) resulting to two quadratic functions: a U-shaped curve and J-shaped curve with the following respective models:

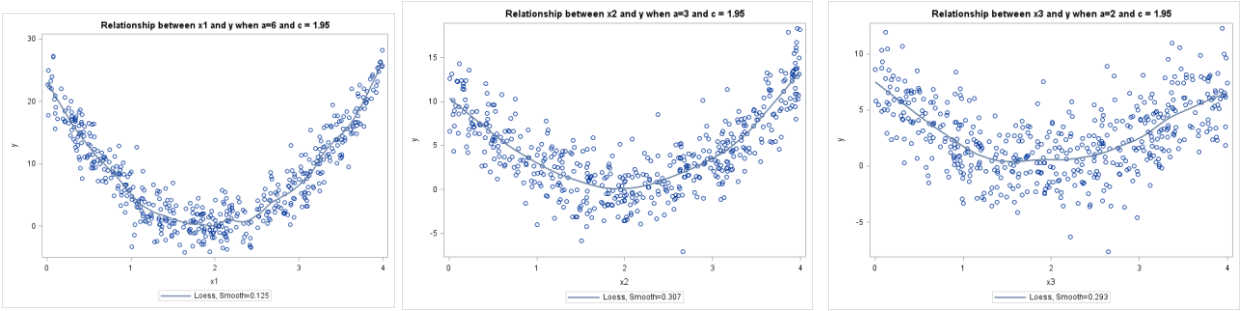
$$Y_i = \delta T_i + \alpha_1 (x_1 - 1.95)^2 + \alpha_2 (x_2 - 1.95)^2 + \alpha_3 (x_3 - 1.95)^2 + \varepsilon_i \qquad (4.5)$$

$$Y_i = \delta T_i + \alpha_1 (x_1 - 1.27)^2 + \alpha_2 (x_2 - 1.27)^2 + \alpha_3 (x_3 - 1.27)^2 + \varepsilon_i \qquad (4.6)$$

The risk status was generated for each subject based on the logit model in Equation (4.1) where we set $\beta_0 = -5.75$ to produce a 50/50 risk and non-risk factor group.

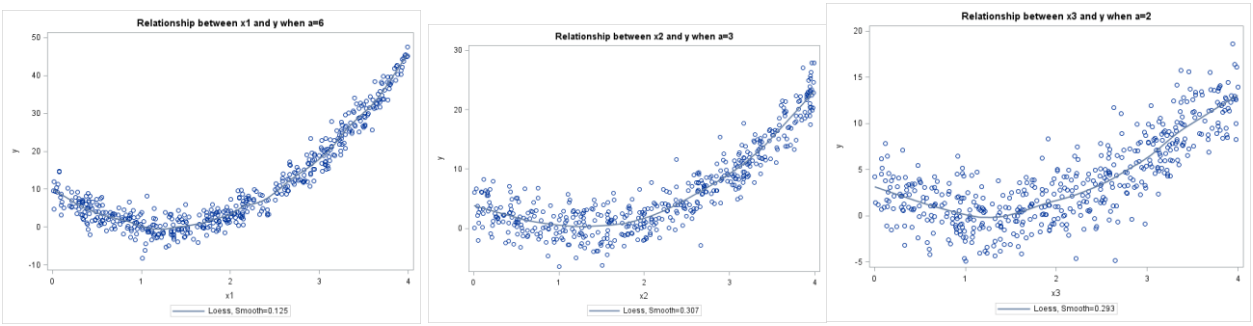
The relationships between the outcome and the covariates based on Equation (4.5) are given in Figure 4.1 and Table 4.4 and those based on Equation (4.6) are given in Figure 4.2 and Table 4.5. From the plots, we can clearly see that Plot A shows a strong relationship between Y and x_1 as the data points fall close to the line. Plots B and C indicate a moderate and a low relationship, respectively. From the tables, x_1 has the highest Hoeffding's correlation, followed by x_2 and then x_3.

Figure 4.1 Plots of association between the outcome and the covariates using Equation (4.5)



Plot A Plot B Plot C

Figure 4.2 Plots of association between the outcome and the covariates using Equation (4.6)



Plot A Plot B Plot C

Table 4.4 Correlation coefficients between the Outcome and the covariates for Model 4.5
Hoeffding Dependence Coefficients

	Y	x1	x2	x3
x1	0.1285	1	0	0
x2	0.0723	0	1	0
x3	0.0436	0	0	1

Table 4.5 Correlation coefficients between the Outcome and the covariates for Model 4.6
Hoeffding Dependence Coefficients

	Y	x1	x2	x3
x1	0.3797	1	0	0
x2	0.2856	0	1	0
x3	0.2140	0	0	1

Under Aim 4, we investigated the influence of modelling a covariate as linear when it is in fact quadratic. We specifically look at the impact of the following models:

a) Modelling x_1 as linear i.e. $Y_i = \delta T_i + \alpha_1 (x_1 - c_1) + \alpha_2 (x_2 - c_2)^2 + \alpha_3 (x_3 - c_3)^2 + \varepsilon_i$

b) Modelling x_2 as linear i.e. $Y_i = \delta T_i + \alpha_1 (x_1 - c_1)^2 + \alpha_2 (x_2 - c_2) + \alpha_3 (x_3 - c_3)^2 + \varepsilon_i$

c) Modelling x_3 as linear i.e. $Y_i = \delta T_i + \alpha_1 (x_1 - c_1)^2 + \alpha_2 (x_2 - c_2)^2 + \alpha_3 (x_3 - c_3) + \varepsilon_i$

d) Modelling x_1, x_2, x_3 as linear i.e. $Y_i = \delta T_i + \alpha_1 (x_1 - c_1) + \alpha_2 (x_2 - c_2) + \alpha_3 (x_3 - c_3) + \varepsilon_i$

4.3 Estimation of Risk effect

The adjusted risk effect was estimated via AUC based on Janes *et al.'s* method (2009) for accommodating covariates in ROC analysis as described in Chapter 2 and briefly in Section 3.6.5.

4.4 Evaluation criteria

Five criteria were used to evaluate the impact of model misspecification in AUC regression adjustment. The first three are bias, relative bias and root mean square error (RMSE) as defined in Sections 3.7.1 and 3.7.2. The other two criteria are the 95% confidence intervals coverage and the type I error. The coverage was used to determine the proportion of times the true mean was contained in the interval estimator. Using the normal approximation interval and our 1,500 simulated data sets, any coverage less than 93.9% and greater than 96.1% is statistically different from 95% as described in Section 3.6.3. The type I error was used to determine the smallest possible error probability in rejecting the true null hypothesis. We fix the level of the test to 0.05.

4.5 Results of the simulation study

Results of the simulation study are given in Tables 4.6 to 4.10 and Figures 4.3 to 4.7. When we examined the impact of missing influential covariates in Aim 1, the results of the mean estimated risk effect across the 1,500 simulated data sets in Table 4.6 and Figure 4.3 showed that the greatest bias is associated with leaving a strong covariate out. The relative bias ranged from 4.2% to 43.72% for AUC = 0.9 to 0.5, respectively. When the model was missing a moderate covariate, the risk estimates were also biased but much less biased than the results with the strong covariate. Furthermore, a model ignoring a covariate weakly related to the outcome did not have a great impact on the estimated risk effect: In all three cases, the bias was almost null; the coverage proportions were significantly less than 95%; and the type I error has the smallest possible error probability of 5.53% . Another important finding is that the RMSE is greater when there is no effect (AUC=0.5) compared to when true AUC is .7 or .9. Finally, modelling a strong covariate alone leads to results similar to the model missing the moderate covariate.

When we examined Aim 2 i.e. what would happen if a continuous covariate is categorized as dichotomous, the results in Table 4.7 and Figure 4.4 suggested that modelling a weak continuous covariate as dichotomous has a superior fit than dichotomizing a moderate and strong covariate. Among the four models in Aim 2, the worst is when all continuous covariates are categorized as dichotomous: the model is associated with the greatest bias and a type I error of 99.93%.

When we investigated the impact of leaving out interaction terms when in reality the true model contains interactions (Aim 3), we found out that any model ignoring any interaction term leads to biased estimates of the true AUC. The greatest biases were associated with models 1, 2,

and 4 when leaving out strong, moderate and all interactions, respectively -See Table 4.8 and Figure 4.5.

Finally, when we examined the effect of non-linearity by modelling a covariate as linear when it is in fact quadratic, we note three important results. First, the U-shaped and the J-shaped models produce the exact same results when the true models are misspecified as seen in Tables 4.9 and 4.10 and in Figures 4.6 and 4.7. Second, when the strong covariate and when all three covariates were considered linear, the true estimates were greatly biased. Finally, to our surprise, AUC of 0.5 is not associated with the greatest bias anymore as we have seen in Aims 1-3 but the risk effect is most biased when AUC = 0.7 or 0.9.

In conclusion, it is far more damaging when misspecification involves a strong covariate than to incorrectly model a covariate weakly associated with the outcome. Among all fitted models from Tables 4.6 to 4.10 and Figures 4.3 to 4.7, the greatest bias is seen when there is no effect (AUC=0.5) except when we incorrectly modelled the non-linear relationship. For the "correct" models or "Model 0" in Tables 4.6 through 4.10 and Figures 4.3 to 4.7, we would expect the models to perform best which is the case. However, the type I error are greater than expected. We speculate that this might be due to the choice of bootstrapping method used in the simulations, or the normality assumption in estimating the parametric AUC might be violated.

Table 4.6 AIM 1 Simulation Results

True AUC	Mean Adjusted AUC	Bias	Relative bias	Bootstrap Standard error	RMSE	95% CI Coverage (%)	Type I Error (%)
Model 0: Real model including all 3 covariates							
0.5	0.4918	-0.0082	-1.6322	0.0184	0.0201	91.60	8.40
0.7	0.6914	-0.0086	-1.2350	0.0166	0.0187	91.60	--
0.9	0.8944	-0.0056	-0.6274	0.0098	0.0113	92.20	--
Model 1: Model missing strong covariate							
0.5	0.7186	0.2186	43.7164	0.0159	0.2192	0.00	100.00
0.7	0.8346	0.1346	19.2353	0.0124	0.1352	0.00	--
0.9	0.9384	0.0384	4.2708	0.0070	0.0391	0.60	--
Model 2: Model missing moderate covariate							
0.5	0.5670	0.0670	13.4021	0.0184	0.0695	5.47	94.53
0.7	0.7446	0.0446	6.3756	0.0157	0.0473	20.47	--
0.9	0.9135	0.0135	1.5045	0.0089	0.0162	62.47	--
Model 3: Model missing weak covariate							
0.5	0.5008	0.0008	0.1503	0.0181	0.0181	94.47	5.53
0.7	0.6976	-0.0024	-0.3403	0.0163	0.0164	94.13	--
0.9	0.8964	-0.0036	-0.3956	0.0095	0.0102	94.13	--
Model 4: Model including strong covariate only							
0.5	0.5769	0.0769	15.3834	0.0181	0.0790	1.27	98.73
0.7	0.7517	0.0517	7.3822	0.0152	0.0539	10.27	--
0.9	0.9163	0.0163	1.8082	0.0086	0.0184	50.40	--

-- Indicates Power = 100%

Figure 4.3 Simulation Results for Aim 1

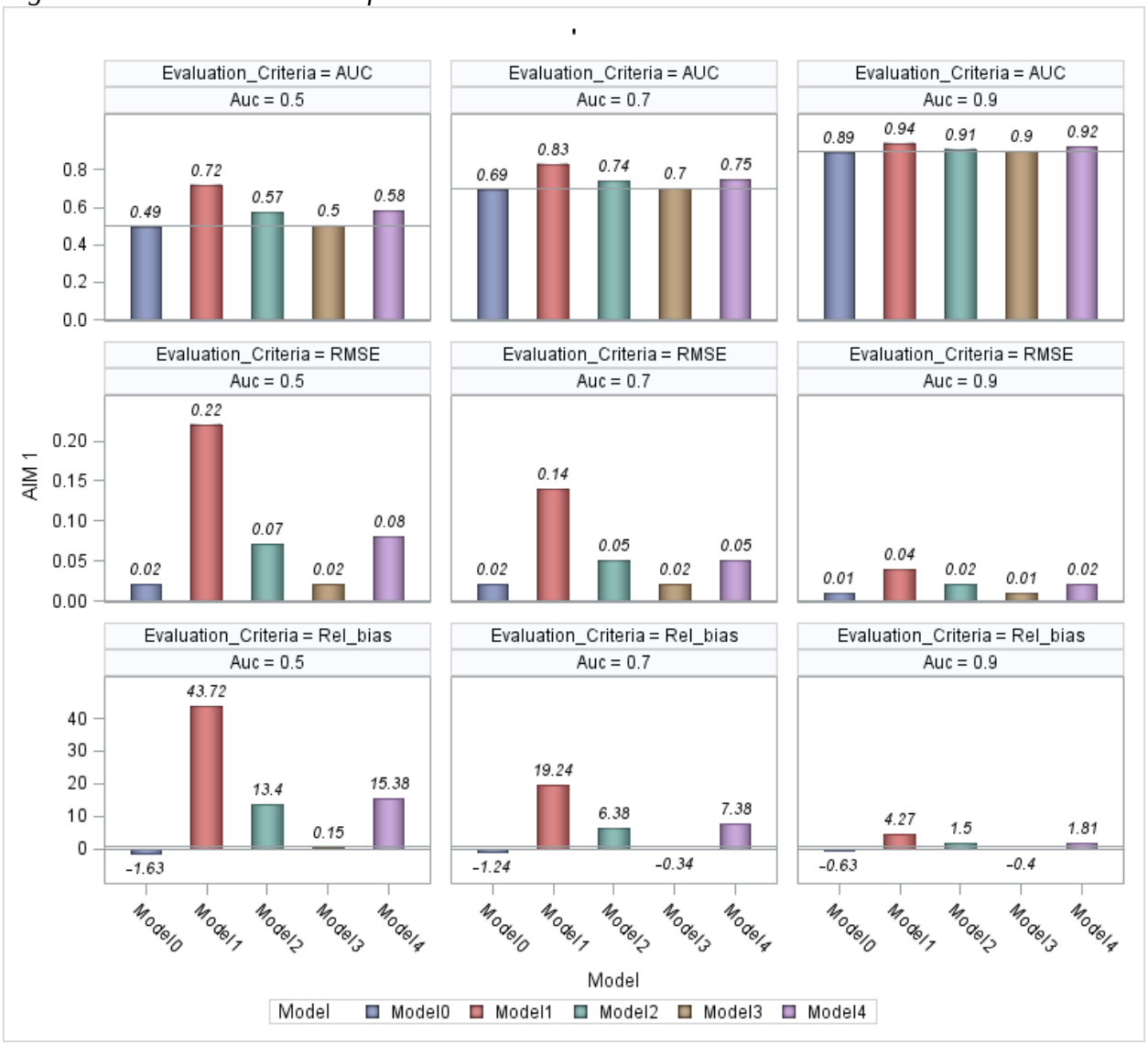


Table 4.7 AIM 2 Simulation Results

True AUC	Mean Adjusted AUC	Bias	Relative bias	Bootstrap Standard error	RMSE	95% CI Coverage (%)	Type I Error (%)
Model 0: Real model including all 3 covariates							
0.5	0.4918	-0.0082	-1.6322	0.0184	0.0201	91.60	8.40
0.7	0.6914	-0.0086	-1.2350	0.0166	0.0187	91.60	--
0.9	0.8944	-0.0056	-0.6274	0.0098	0.0113	92.20	--
Model 1:Covariate strongly associated with the outcome is dichotomized							
0.5	0.5879	0.0879	17.5782	0.0180	0.0897	0.40	99.60
0.7	0.7500	0.0500	7.1442	0.0152	0.0523	12.47	--
0.9	0.9082	0.0081	0.9052	0.0089	0.0121	79.07	--
Model 2: Covariate moderately associated with the outcome is dichotomized							
0.5	0.5090	0.0090	1.8008	0.0186	0.0207	91.67	8.33
0.7	0.7035	0.0035	0.5041	0.0166	0.0170	93.53	--
0.9	0.8986	-0.0014	-0.1603	0.0097	0.0098	93.80	--
Model 3: Covariate weakly associated with the outcome is dichotomized							
0.5	0.4959	-0.0041	-0.8291	0.0183	0.0187	93.53	6.47
0.7	0.6942	-0.0058	-0.8239	0.0165	0.0174	93.33	--
0.9	0.8954	-0.0046	-0.5147	0.0097	0.0107	92.80	--
Model 4: All three covariates are dichotomized							
0.5	0.6053	0.1053	21.0678	0.0178	0.1068	0.07	99.93
0.7	0.7626	0.0626	8.9433	0.0149	0.0643	3.87	--
0.9	0.9135	0.0135	1.4996	0.0086	0.0160	61.73	--

-- Indicates Power = 100%

Figure 4.4 Simulation Results for Aim 2

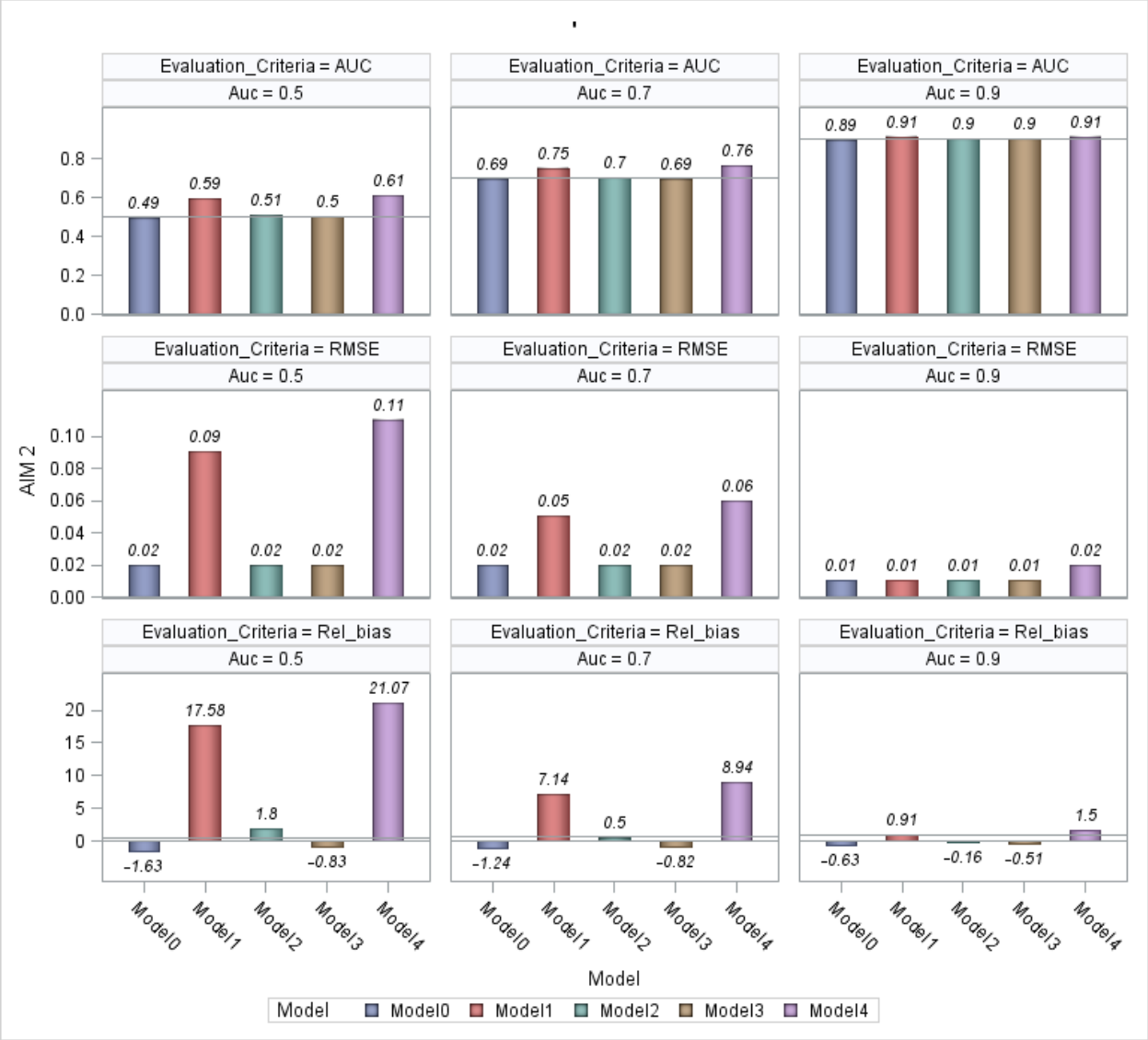


Table 4.8 AIM 3 Simulation Results

True AUC	Mean Adjusted AUC	Bias	relative bias	Bootstrap Standard error	RMSE	95% CI Coverage (%)	Type I Error (%)
Model 0: Real model including all 3 interactions							
0.5	0.4898	-0.0102	-2.0388	0.0185	0.0211	90.53	9.47
0.7	0.6883	-0.0117	-1.6684	0.0168	0.0204	89.40	--
0.9	0.8919	-0.0081	-0.9006	0.0100	0.0129	89.33	--
Model 1: Model ignoring strong interactions							
0.5	0.6589	0.1589	31.7811	0.0185	0.1600	0.00	100.00
0.7	0.7496	0.0495	7.0784	0.0165	0.0522	16.33	--
0.9	0.8534	-0.0466	-5.1764	0.0131	0.0484	3.00	--
Model 2: Model ignoring moderate interactions							
0.5	0.6527	0.1527	30.5472	0.0187	0.1539	0.00	100.00
0.7	0.7458	0.0458	6.5404	0.0168	0.0488	22.40	--
0.9	0.8520	-0.0480	-5.3300	0.0134	0.0498	2.27	--
Model 3:Model ignoring weak interactions							
0.5	0.5154	0.0154	3.0772	0.0187	0.0242	86.60	13.40
0.7	0.6985	-0.0015	-0.2163	0.0168	0.0169	94.73	--
0.9	0.8875	-0.0125	-1.3931	0.0105	0.0163	80.07	--
Model 4: Model ignoring all interactions							
0.5	0.6591	0.1591	31.8262	0.0185	0.1602	0.00	100.00
0.7	0.7499	0.0499	7.1252	0.0165	0.0525	16.00	--
0.9	0.8538	-0.0462	-5.1389	0.0131	0.0481	3.47	--

-- Indicates Power = 100%

Figure 4.5 Simulation Results for Aim 3

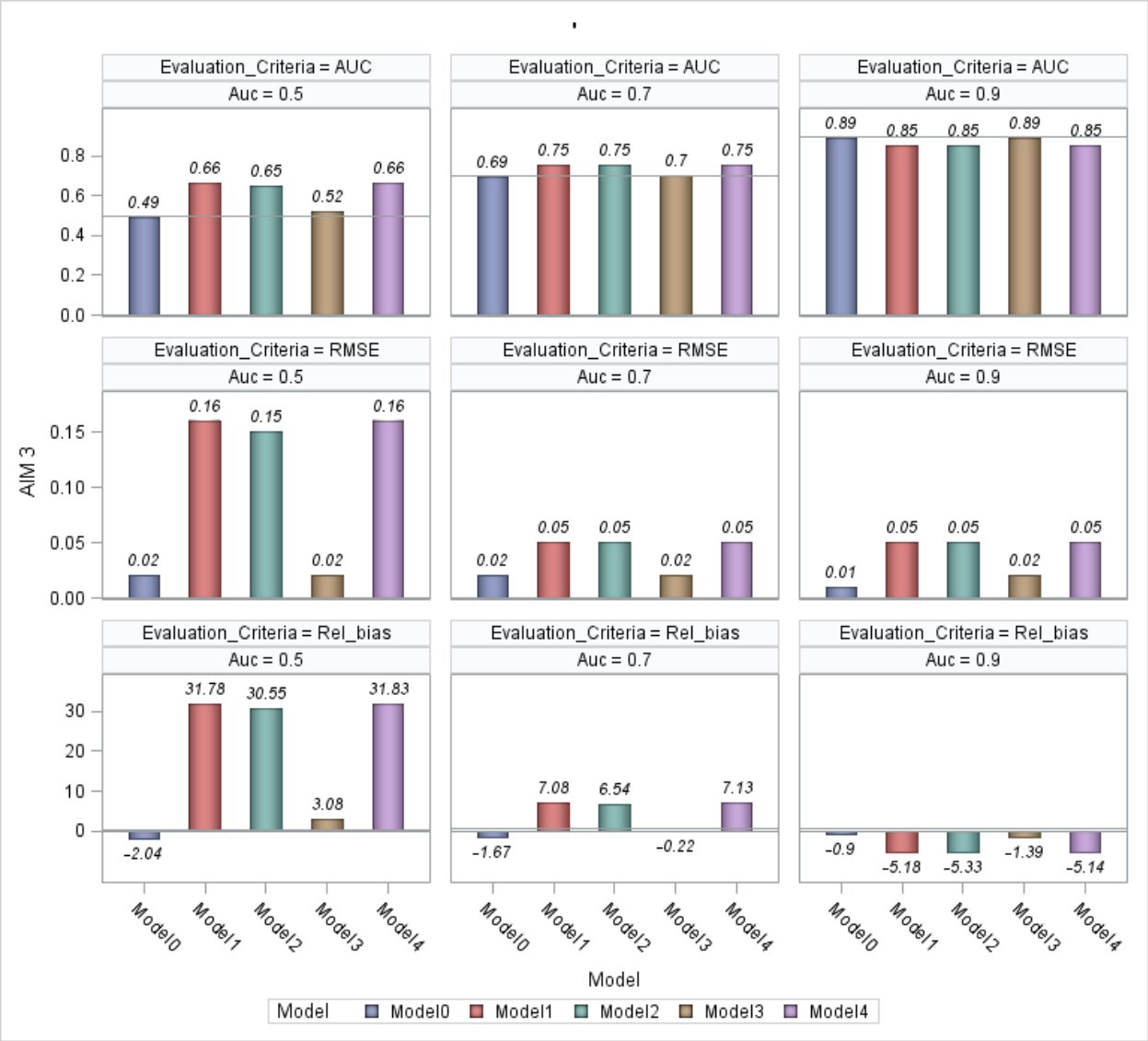


Table 4.9 AIM 4 U-shaped Simulation Results

True AUC	Mean Adjusted AUC	Bias	relative bias	Bootstrap Standard error	RMSE	95% CI Coverage (%)	Type I Error (%)
Model 0: Real model including all covariates							
0.5	0.5560	0.0560	11.2050	0.0181	0.0589	86.13	13.87
0.7	0.7433	0.0433	6.1812	0.0153	0.0459	21.87	--
0.9	0.9183	0.0183	2.0316	0.0082	0.0200	40.93	--
Model 1: Modelling x1 as linear							
0.5	0.6567	0.1567	31.3338	0.0205	0.1580	0.00	100.00
0.7	0.7156	0.0156	2.2351	0.0184	0.0242	84.53	--
0.9	0.7960	-0.1040	-11.5545	0.0148	0.1050	0.00	--
Model 2: Modelling x2 as linear							
0.5	0.5553	0.0553	11.0682	0.0187	0.0584	16.33	83.67
0.7	0.6692	-0.0308	-4.4040	0.0168	0.0351	55.00	--
0.9	0.8112	-0.0888	-9.8670	0.0124	0.0897	0.00	--
Model 3: Modelling x3 as linear							
0.5	0.5336	0.0336	6.7136	0.0184	0.0383	54.60	45.40
0.7	0.6743	-0.0257	-3.6707	0.0167	0.0306	66.60	--
0.9	0.8392	-0.0608	-6.7527	0.0117	0.0619	0.07	--
Model 4: Modelling all 3 covariates as linear							
0.5	0.6751	0.1751	35.0189	0.0195	0.1762	0.00	100.00
0.7	0.7244	0.0244	3.4901	0.0179	0.0303	69.40	--
0.9	0.7906	-0.1094	-12.1606	0.0153	0.1105	0.00	--

-- Indicates Power = 100%

Figure 4.6 Simulation Results for U-shaped Aim 4

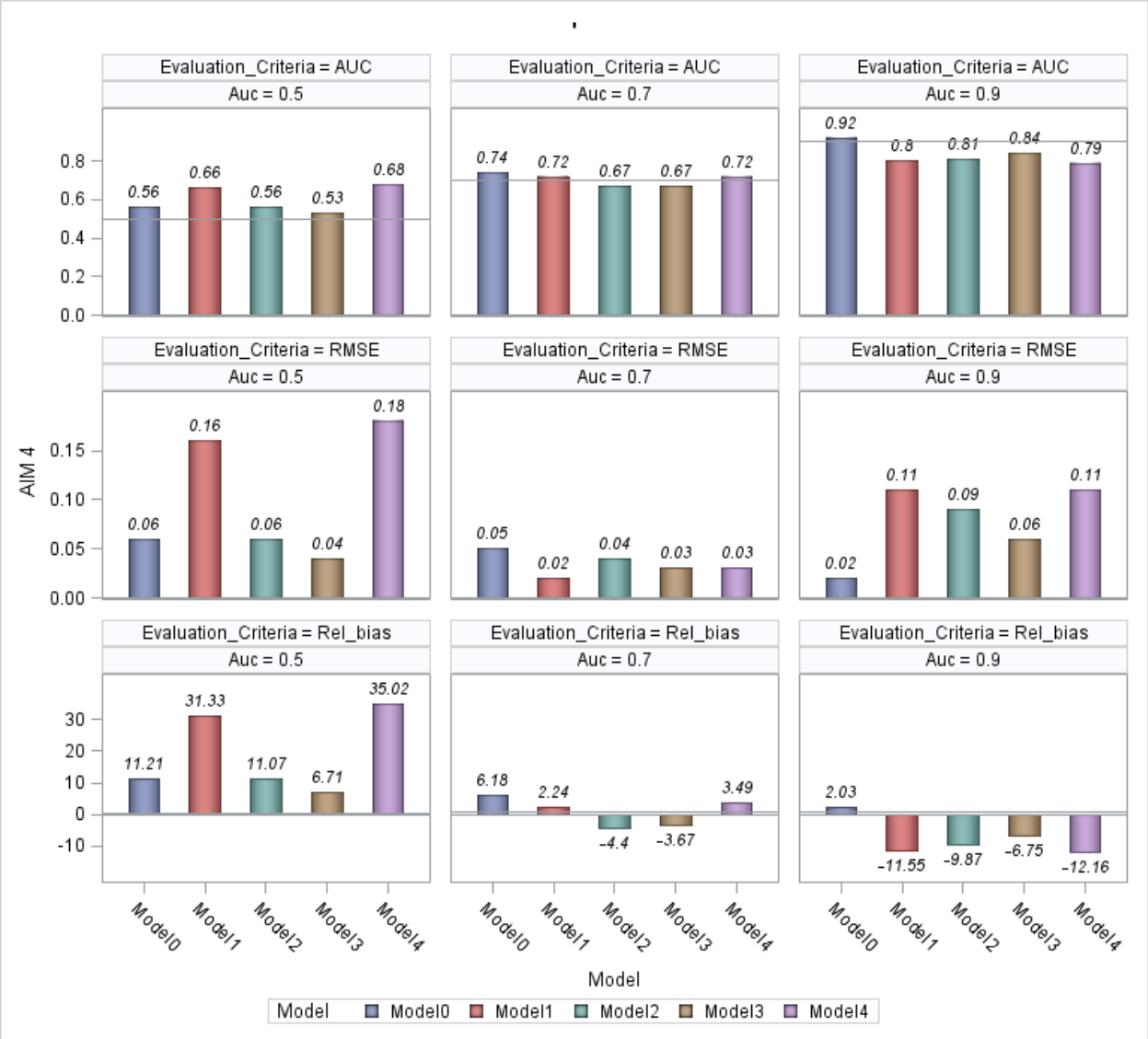
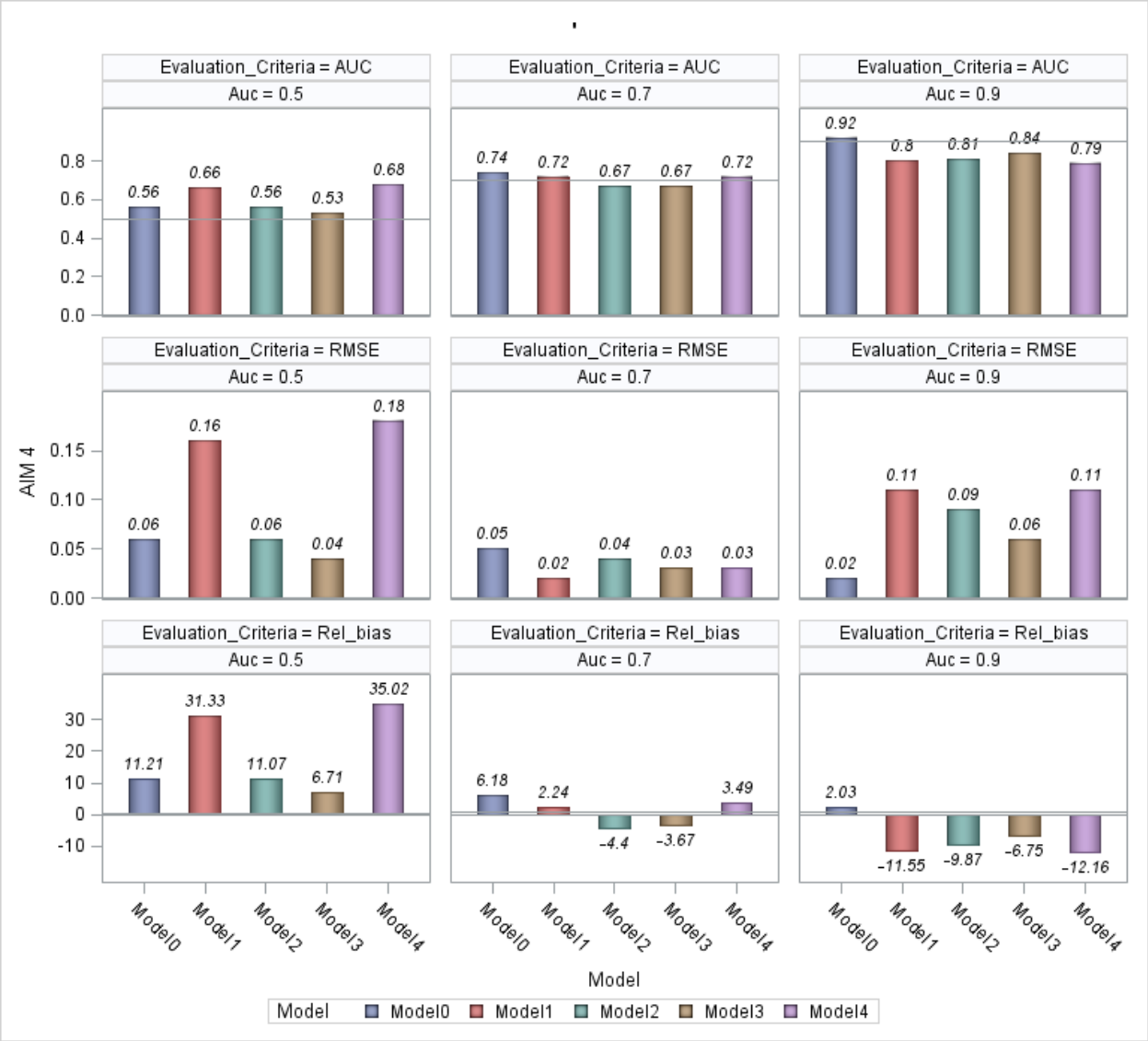


Table 4.10 AIM 4 J-shaped Simulation Results

True AUC	Mean Adjusted AUC	Bias	relative bias	Bootstrap Standard error	RMSE	95% CI Coverage (%)	Type I Error (%)
Model 0: Real model including all covariates							
0.5	0.5560	0.0560	11.2050	0.0181	0.0589	86.13	13.87
0.7	0.7433	0.0433	6.1812	0.0153	0.0459	21.87	--
0.9	0.9183	0.0183	2.0316	0.0082	0.0200	40.93	--
Model 1: Modelling x1 as linear							
0.5	0.6567	0.1567	31.3338	0.0205	0.1580	0.00	100.00
0.7	0.7156	0.0156	2.2351	0.0184	0.0242	84.53	--
0.9	0.7960	-0.1040	-11.5545	0.0148	0.1050	0.00	--
Model 2: Modelling x2 as linear							
0.5	0.5553	0.0553	11.0682	0.0187	0.0584	16.33	83.67
0.7	0.6692	-0.0308	-4.4040	0.0168	0.0351	55.00	--
0.9	0.8112	-0.0888	-9.8670	0.0124	0.0897	0.00	--
Model 3: Modelling x3 as linear							
0.5	0.5336	0.0336	6.7136	0.0184	0.0383	54.60	45.40
0.7	0.6743	-0.0257	-3.6707	0.0167	0.0306	66.60	--
0.9	0.8392	-0.0608	-6.7527	0.0117	0.0619	0.07	--
Model 4: Modelling all 3 covariates as linear							
0.5	0.6751	0.1751	35.0189	0.0195	0.1762	0.00	100.00
0.7	0.7244	0.0244	3.4901	0.0179	0.0303	69.40	--
0.9	0.7906	-0.1094	-12.1606	0.0153	0.1105	0.00	--

-- Indicates Power = 100%

Figure 4.7 Simulation Results for J-shaped Aim 4



CHAPTER 5: APPLICATION

In the first part of this chapter, we apply different propensity score methods and models to data from the Shock Research Unit at the University of Southern California, Los Angeles, California. We also compare the propensity score results to corresponding results derived from the AUC direct regression for estimating the adjusted AUC.

In the second part of the chapter, we examine model misspecification in AUC regression by incorrectly modelling the covariates in the data.

5.1 Introduction

In the United States, there are more than 1 million admissions to emergency departments annually due to shock, according to Merck (Merck, 2009). The medical disorder of shock is mostly characterized by an abnormal low systolic blood pressure or hypotension. For people in shock, the tissues of the body do not receive enough blood. As a result, the tissues with impaired circulation suffer damage from lack of oxygen. Damage to tissues and organs of the body can lead to severe disability or death of patients in shock.

In this data analysis, it is of primary interest to compare patients in shock with patients not in shock upon admission to the shock research unit at the University of Southern California in Los Angeles, California, in terms of diastolic blood pressure (DBP) at discharge. In other words, we sought to study whether shock status at admission is a risk factor for diastolic blood pressure at discharge. Because this is an observational study where subjects are assigned to one of the risk factor groups in a non-random manner, there is a possibility that discharge diastolic blood pressure in subjects with shock and non-shock is related to some baseline covariates such

as demographic or physiological variables. In order to accurately estimate the risk effect, it is of practical importance to account for those covariates. Failing to adjust for those variables could lead to biased estimates of shock status effect on diastolic blood pressure at discharge.

Therefore, we illustrate our dissertation research findings by using different propensity score methods and models to estimate the probability that the DBP response from the j^{th} randomly chosen patient in the shock group is less than that from the i^{th} randomly selected patient in the non-shock group. This is defined as $P(Y_{NS} > Y_S)$ where Y_S and Y_{NS} are diastolic blood pressure measure for patients in shock and non-shock groups, respectively.

5.2 Methods

The data consists of 113 critically ill patients who were admitted to the Shock Research Unit at the University of Southern California, Los Angeles, California. Data on many physiological variables were collected successively in time on each patient. From the wealth of data that was collected, the present data is a subset that appeared in the book "Statistical Analysis: A computer Oriented Approach" by Afifi and Azen in 1979 for examples and exercises purposes (Afifi & Azen, 1979). In this set, initial measurements (that is, measurements upon admission) and final measurements on the same variables (that is, measurements just before death or discharge) were collected. Hence, each patient has 2 records and each record contains 21 variables: 6 general variables and 14 physiological variables. The outcome of interest was diastolic blood pressure at discharge, whereas the risk factor of interest was shock. A patient is defined as having a shock if experiencing any of these shocks: hypovolemic shock, cardiogenic shock, bacterial shock, neurogenic shock or other.

For each subject, propensity scores (PS) were estimated by fitting a logistic regression to predict shock, as a function of baseline covariates. We constructed four different PS models including different combinations of measured covariates: PS model 1 (PS-M1) included variables associated with the shock status group. The association between the covariates and the shock group were determined using a t-test for continuous covariates and a chi-square test for categorical variables at 5% significance level. PS model 2 (PS-M2) included variables associated with the outcome, diastolic blood pressure. The association between the outcome and the continuous covariates were measured using a Pearson correlation; and a t test is used to test association between the outcome and the categorical covariates. PS model 3 (PS-M3) included variables associated with both the risk factor group and the outcome i.e. all common covariates to the previous two models. PS model 4 (PS-M4) included all measured variables.

As described in Section 3.4, strata were created based on the quintiles of the estimated propensity scores; and also matched pairs of subjects in shock and not in shock were created using the 1:1 greedy matching technique with calipers of width 0.2 of the standard deviation of the logit of the propensity score.

Two analyses were carried out: The first analysis comprised all measured variables while the second analysis included only uncorrelated variables. The reason for the second analysis was to more closely mimic the simulations, which assumed uncorrelated variables. Four different methods were used to estimate the adjusted AUC in diastolic blood pressure at discharge. First, subjects were stratified based on the quintiles of the propensity score and the adjusted AUC was computed as described in section 3.5.2. Second, we estimated the adjusted shock effect via AUC in the propensity score matched sample as described in Section 3.5.3. Third, the risk group effect is estimated under the covariate adjustment on the propensity score method using the method

described in 3.5.4. Finally, for comparison purposes, we used the direct AUC regression method to adjust for covariates in directly modelling covariates effects on the response as described in 3.5.5. For this method, we considered 4 separate regression models as well where each model is described above.

Furthermore, the issue of model misspecification in AUC regression method was investigated. Eight different models were used: these models included 4 models missing influential variables and 4 models where continuous variables were modelled as dichotomous. The estimated AUCs were computed and then compared to the adjusted AUC obtained from our "best" model.

5.3 Results

The study sample consisted of n=113 critically ill subjects of whom 79 (69.91%) where in shock and 34 (30.09%) were not in shock upon admission to the shock research unit. Table 5.1 shows the summary statistics of the baseline covariates between subjects in shock and not in shock. Subjects in the two groups were compared using pooled t-tests and chi-square tests for continuous and dichotomous variables, respectively. The descriptive analysis reveals that patients in shock have lower values of mean arterial pressure, systolic and diastolic pressure upon admission ($p<0.0001$, $p<0.0001$ and $p = 0.0042$, respectively). Heart rate beats and mean circulation time tend to be higher in those in shock than those not in shock (all p-value <0.05). Surprisingly, the urinary output is almost three times higher in patients not in shock compared to patients in shock ($p=0.0019$). There was no statistical difference between the two groups of patients in regards to age, mean central venous pressure, body surface index, appearance time, red cell index, hemoglobin, hematocrit and gender.

Table 5.1 Baseline Characteristics of the Study Sample by Shock Group

Variable		Shock	No-shock	Total	P-value	Pearson r
		N = 79 (69.91%)	N =34 (30.09%)	N=113		
Age (years)		55.1 ± 16.9	53.4 ± 15.9	54.6 ± 16.6	0.6114	-0.0483
Systolic Pressure (mm Hg)		97.8 ± 28.9	127.6 ± 23.4	106.2 ± 30.7	<.0001	0.4577
Mean arterial pressure (mm Hg)		68.2 ± 21.6	85.5 ± 18.1	73.4 ± 22	<.0001	0.3625
Heart rate (beats/min)		108.8 ± 28.1	94.3 ± 31.0	104.4 ± 29.6	0.0158	-0.2265
Diastolic pressure (mm Hg)		55.3 ± 18.8	66.1 ± 15.7	58.5 ± 18.5	0.0042	0.2676
Mean central venous pressure (cmH2O)		9.1 ± 5.5	8.4 ± 6.2	8.9 ± 5.7	0.5272	-0.0601
Body surface index (m2)		1.7 ± 0.2	1.7 ± 0.8	1.7 ± 0.2	0.4165	0.0772
Cardiac index (liters/min m2)		2.36 ± 1.4	3.1 ± 1.5	2.6 ± 1.5	0.0190	0.2203
Appearance time (sec)		10.78 ± 5.0	8.9 ± 4.3	10.2 ± 4.9	0.0570	-0.1796
Mean circulation time (sec)		24.1 ± 10.8	19.7 ± 9.0	22.8 ± 10.5	0.0391	-0.1944
Urinary output (ml/hr)		33.2 ± 79.0	103.7 ± 156.6	54.4 ± 112.3	0.0019	0.2888
Plasma volume index (ml/kg)		47.0 ± 15.7	52.9 ± 13.5	48.8 ± 15.2	0.0579	0.1789
Red Cell Index (ml/kg)		21.1 ± 9.4	22.0 ± 7.1	21.4 ± 8.7	0.6444	0.0439
Hemoglobin (gm/100 ml)		11.6 ± 2.6	11.0 ± 2.2	11.4 ± 2.5	0.2834	-0.1018
Hematocrit (percent)		35.3 ± 8.1	33.9 ± 7.1	34.9 ± 7.8	0.3687	-0.0854
Sex	Male	38 (48.10 %)	21 (61.76 %)	59 (52.21%)	-0.1255	-0.1255
	Female	41 (51.90 %)	13 (38.24 %)	54 (47.79%)		

Continuous variables are reported as mean ± standard deviation. Dichotomous variables are reported as frequency and percent.

The selection of the variables entering the different PS models is summarized in Table 5.2. The number of covariates in the PS models range from 3 to 16. Logistic regression is used to estimate the propensity scores and the c-statistic, a measure known to measure model fit for logistic regression is also reported. The unadjusted model doesn't contain any variable; it has a c-statistic value of 0.7174. The PS-M1 contains 7 covariates with a c-statistic of 0.85; PS-M2 has 6 variables and a c-statistic of 0.826; PS-M3 contains 3 covariates with c-statistic of 0.812; and finally PS-M4 contains all 16 variables and has a c-statistic of 0.895.

The crude AUC between shock groups was 0.7174. We obtain a 95% confidence interval for the unadjusted AUC of (0.61-0.82) using the Delong formula incorporated into SAS PROC LOGISTIC. Since the confidence interval does not contain the null value 0.5, we conclude that for two randomly chosen patients from the shock and non-shock groups, the probability that the DBP response from the participant in the non-shock group exceeds that the response from the patient in the shock group is estimated to be 71.74%. In other words, there is a significant chance that the DBP of those in the non-shock group is greater than that of those in the shock group.

Table 5.2 Selection of variables entering different propensity score models

Covariates	PS model 1	PS model 2	PS model 3	PS model 4
Age (years)				✓
Systolic Pressure (mm Hg)	✓	✓	✓	✓
Mean arterial pressure (mm Hg)	✓	✓	✓	✓
Heart rate (beats/min)	✓			✓
Diastolic pressure (mm Hg)	✓	✓	✓	✓
Mean central venous pressure (cmH2O)		✓		✓
Body surface index (m2)		✓		✓
Cardiac index (liters/min m2)	✓			✓
Appearance time (sec)				✓
Mean circulation time (sec)	✓			✓
Urinary output (ml/hr)	✓			✓
Plasma volume index (ml/kg)		✓		✓
Red Cell Index (ml/kg)				✓
Hemoglobin (gm/100 ml)				✓
Hematocrit (percent)				✓
Sex				✓

The adjusted estimates using four different methods are reported in Table 5.3. Using stratification on the quintiles of the propensity score, the adjusted estimates of $P\left(Y_{NS} > Y_S\right)$ range from 0.6337 to 0.6833 for different PS models. The standard error and the confidence interval were obtained using Equation 3-18. In contrast to the unadjusted AUC, all four 95 per cent confidence intervals contain the null value of 0.5. This indicates that under stratification, the adjusted AUC is not statistically different from the null value i.e. we fail to reject $H_o : AUC = 0.5$.

Propensity score matching resulted in the formation of 22, 26, 28 and 21 pairs of subjects in shock and not in shock for PS models 1, 2, 3, and 4 respectively. The adjusted estimates range from 0.599 to 0.675. The standard error and the 95% confidence interval were obtained using 1000 bootstrap samples of the original observations. The results are not consistent under this method as some confidence intervals contain the null value and some do not.

Using covariate adjustment on the propensity score, inconsistency of the results is similar to what we found with matching. The standard errors were also estimated based on 1000 bootstraps.

The fourth method consisting of using the AUC regression to directly model the covariates on the response resulted in values of adjusted AUC close to each other ranging from 0.596 to 0.639. All four confidence intervals estimated based on bootstrap are consistent and contain the null value of 0.5.

From the results of the Monte Carlo simulations described in Section 3.7, we found the crude estimate is biased positively when true AUC is around 0.5 and 0.7. Hence, the crude estimate in this application is most likely overestimated. Furthermore, we found out that stratifying, matching and covariate adjustment on the propensity score resulted in biased estimation of AUC. Thus, in our illustration study, these estimates obtained from the propensity score methods and models are likely subject to a great deal of bias. Given the results and the recommendations of our simulations, our best estimate of the true risk effect is most likely the estimate obtained from model 2 using the direct AUC regression adjustment. Hence the adjusted estimate of the probability that the DBP response from the j^{th} randomly chosen patient in the shock group is less than that from the ith randomly patient in the non-shock group i.e. $P\left(Y_{NS} > Y_S\right)$ is 0.6224; the 95% CI based on 1,000 bootstrap samples is (0.4633, 0.7919). Since

the confidence interval includes 0.5, there is no statistical evidence that shock status upon admission is a risk factor for diastolic blood pressure.

Table 5.3 Effect estimates from different methods and models

Models/Methods	AUC	SE	95%CI
Unadjusted	0.7174	0.0529	0.6138 - 0.8210
PS Stratify -M1	0.6734	0.1478	0.3837 - 0.9630
PS Stratify -M2	0.6404	0.1239	0.3975 - 0.8833
PS Stratify -M3	0.6833	0.1209	0.4463 - 0.9204
PS Stratify -M4	0.6337	0.1599	0.3203 - 0.9471
PS Matching -M1	0.59902	0.0867	0.4330 - 0.7728
PS Matching -M2	0.67498	0.07423	0.5314 - 0.8224
PS Matching -M3	0.6698	0.0700	0.5349 - 0.8094
PS Matching -M4	0.64918	0.0825	0.4888 - 0.8120
PS Covariate Adjust - M1	0.62574	0.08133	0.4712 - 0.7900
PS Covariate Adjust - M2	0.64133	0.06598	0.5128 - 0.7714
PS Covariate Adjust - M3	0.64756	0.0629	0.5265 - 0.7731
PS Covariate Adjust - M4	0.6117	0.07937	0.4550 - 0.7661
Reg. Adjustment - M1	0.60579	0.08173	0.4467 - 0.7670
Reg. Adjustment - M2	0.62241	0.08382	0.4633 - 0.7919
Reg. Adjustment - M3	0.63894	0.07523	0.4917 - 0.7865
Reg. Adjustment - M4	0.59619	0.09238	0.4208 - 0.7829

In the second part of our data illustration, we aim to explore model misspecification. The "best" model is considered as the model containing the 6 variables associated with the outcome (Regression adjustment Model 2). Each of these variables varies in their association with the response DBP according to the Pearson correlations values and their classification in Table 5.4.

Table 5.4 Classification of baselines covariates based on their association with outcome

Variable	Pearson r	Strong Covariates	Moderate Covariates	Weak Covariates
Age (years)	-0.01490			
Systolic Pressure (mm Hg)	0.47679	✓		
Mean arterial pressure (mm Hg)	0.52209	✓		
Heart rate (beats/min)	0.03178		✓	
Diastolic pressure (mm Hg)	0.52558		✓	
Mean central venous pressure (cmH2O)	-0.29073			
Body surface index (m2)	0.31300			
Cardiac index (liters/min m2)	0.04843		✓	
Appearance time (sec)	-0.10126			
Mean circulation time (sec)	-0.13453			✓
Urinary output (ml/hr)	0.15508		✓	
Plasma volume index (ml/kg)	-0.19066			
Red Cell Index (ml/kg)	0.03015			
Hemoglobin (gm/100 ml)	0.12164			
Hematocrit (percent)	0.13331			
Sex -0.12310 Female	-0.12550			

Hence, systolic, diastolic and mean arterial pressure are strongly related to outcome (r ranging from 0.48 to 0.53 and p-values are highly significant, $p<0.0001$). Mean central venous pressure and body surface index are moderately associated with DBP: r value close to 0.3. Finally, plasma volume index is weakly related to DBP, r = -0.19 and p-value = 0.0431.

The estimates of $P\left(Y_{NS} > Y_S\right)$ in model misspecification are given in Table 5.5. When we examined the impact of missing influential covariates, the greatest bias was associated with

leaving the strong covariates out. The estimated AUC was 0.736 with a 95% CI of (0.6253, 0.8436). There's not much harm in leaving a weak covariate out. These results were consistent with our findings from the Monte Carlo simulations.

From our simulations results in Section 4.5, we found that modelling a weak covariate as dichotomous has a superior fit than modelling a strong covariate as dichotomous. The findings in our case study are consistent with those of the simulations study. The worst model in dichotomization is modelling all covariates as dichotomous when they are in fact continuous.

From the case study, we conclude that it is far more damaging to incorrectly model covariates strongly associated with the outcome than to incorrectly model covariates weakly associated with the outcome. These findings were consistent with previous findings in our Monte Carlo simulations.

Table 5.5 Effect Estimates from Model Misspecification

Models	**AUC**	**SE**	**95%CI**
Best model	0.6224	0.0838	0.4633 - 0.7919
Missing strong covariates	0.7360	0.0557	0.6253 - 0.8436
Missing moderate covariates	0.6463	0.0798	0.4917 - 0.8045
Missing weak covariates	0.6168	0.0811	0.4627 - 0.7806
Including strong covariates only	0.6389	0.0752	0.4917 - 0.7865
Strong covariates are dichotomized	0.7083	0.0651	0.5815 - 0.8367
Moderate covariates are dichotomized	0.6262	0.0817	0.4681 - 0.7883
Weak covariates are dichotomized	0.6182	0.0843	0.4595 - 0.7899
All covariates are dichotomized	0.7074	0.0694	0.5719 - 0.8441

For our second analysis, a subset of covariates not correlated with each other was selected using the Pearson correlation criteria. A total of 6 variables was considered as compared to 16 variables in the previous analysis – See Table 5.6-.

Table 5.6 Pearson Correlation Coefficients, N = 113

	MAP	HR	MCVP	CI	BSI	RCI
Mean arterial pressure in mm Hg	1	-0.0702	-0.0778	0.04001	0.21098	0.04198
Heart rate in beats/min	-0.0702	1	0.05307	-0.0296	-0.0464	-0.0398
Mean central venous pressure in cmH2O	-0.0778	0.05307	1	0.00248	0.0763	-0.0566
Cardiac index in liters/min m2	0.04001	-0.0296	0.00248	1	0.0494	-0.1206
Body surface index in m2	0.21098	-0.0464	0.0763	0.0494	1	-0.0462
Red Cell Index in ml/kg	0.04198	-0.0398	-0.0566	-0.1206	-0.0462	1

The selection of the variables entering the different PS models is summarized in Table 5.7. The number of covariates in the PS models range from 1 to 6. Hence, the PS-M1 contains 3 covariates; PS-M2 has 3; PS-M3 contains 1 covariate; and finally PS-M4 contains all 6 uncorrelated variables.

Table 5.7 Selection of variables entering different propensity score models

Covariates	**PS model 1**	**PS model 2**	**PS model 3**	**PS model 4**
Mean arterial pressure (mm Hg)	✓	✓	✓	✓
Heart rate (beats/min)	✓			✓
Mean central venous pressure (cmH2O)		✓		✓
Body surface index (m2)		✓		✓
Cardiac index (liters/min m2)	✓			✓
Red Cell Index (ml/kg)				✓

The adjusted AUC estimates were given in Table 5.8. Stratifying on the quintiles of the propensity score yield to adjusted AUCs ranging from 0.62 to 0.67. Under the stratification method, all four 95% confidence intervals contain the null value of 0.5 which indicate that there is no significant chance that the DBP of those in the non-shock group is greater than that of those in the shock group.

Table 5.8 Effect estimates from different methods and models

Models/Methods	**AUC**	**SE**	**95%CI**
Unadjusted	0.7174	0.0529	0.6138 - 0.8210
PS Stratify -M1	0.6223	0.136	0.3557 - 0.8889
PS Stratify -M2	0.6717	0.1103	0.4555 - 0.8879
PS Stratify -M3	0.6573	0.1397	0.4485 - 0.8661
PS Stratify -M4	0.6300	0.1397	0.3561 - 0.9039
PS Matching -M1	0.6209	0.0810	0.4669 -0.7846
PS Matching -M2	0.6196	0.0725	0.4778 - 0.7619
PS Matching -M3	0.6206	0.0722	0.4779 - 0.7608
PS Matching -M4	0.6768	0.0744	0.5339 - 0.8257
PS Covariate Adjust - M1	0.6301	0.0675	0.5013 - 0.7661
PS Covariate Adjust - M2	0.6693	0.0535	0.5668 - 0.7764
PS Covariate Adjust - M3	0.6717	0.0540	0.5684 - 0.7799
PS Covariate Adjust - M4	0.6315	0.0697	0.4999 - 0.7730
Reg. Adjustment - M1	0.6685	0.0625	0.5471- 0.7919
Reg. Adjustment - M2	0.6770	0.0557	0.5697 - 0.7880
Reg. Adjustment - M3	0.6661	0.0521	0.5643 - 0.7686
Reg. Adjustment - M4	0.6831	0.0668	0.5532 - 0.8149

For propensity score matching and covariate adjustment on the propensity score methods, the adjusted AUCs range from 0.62 to 0.68. The standard error and the 95% confidence interval were obtained using 1000 bootstrap samples of the original observations. The results are not consistent under these two methods as some confidence intervals contain the null value and some do not. For AUC regression adjustment, the values of the adjusted AUCs are very close to each other and close to 0.67. All confidence intervals are consistent and do not contain the null value of 0.5. Hence, we conclude that there is a significant probability that DBP when in shock status is greater than DBP when in no-shock status.

The results obtained from the propensity score methods in this second analysis are consistent with those obtained from the first analysis. However, with the regression adjustment method, it appears that correlation between variables has a great effect on the estimate of the

adjusted AUC: The AUC estimates were not significant different from 0.5 when the covariates were correlated but they were significant when the covariates were uncorrelated. Given these findings, correlation between covariates should be taken into account when estimating AUC through regression adjustment. One may also speculate that the difference between the two analyses might be due to the fact that the first analysis included more covariates (up to 16) while the second analysis included only a maximum of 6 covariates and we know that excluding covariates can lead to incorrect answers.

In the second part of this second analysis, we explore model misspecification. From the recommendations of the simulations study, we consider the "best" model to be model 2 for Regression Adjustment in Table 5.8. This model refers to the model containing the 3 variables associated with the outcome. Each of these three variables varies in their association with the response DBP according to the correlations values and their classification in Table 5.4. Therefore, mean arterial pressure is strongly related to outcome (r=0.52) while mean central venous pressure and body surface index are moderately associated with DBP (r ~ 0.3). There was no variable weakly associated with the outcome in this subset analysis. The AUC estimates from model misspecification are given in Table 5.9. We notice that the greatest harm is associated with leaving a strong covariate out as compared to leaving a moderate covariate out. These results were consistent with our findings from the Monte Carlo simulations.

Table 5.9 Effect Estimates from Model Misspecification

Models	AUC	SE	95%CI
Best model	0.67699	0.0557	0.5697 - 0.7880
Missing strong covariates	0.721	0.053	0.616- 0.824
Missing moderate covariates	0.666	0.052	0.564 - 0.769
Including strong covariates only	0.666	0.052	0.564- 0.769
Strong covariates are dichotomized	0.694	0.055	0.586 - 0.803
Moderate covariates are dichotomized	0.666	0.054	0.561 - 0.772
All covariates are dichotomized	0.687	0.054	0.582 - 0.791

Furthermore, we investigated the effect of dichotomizing continuous variables in model misspecification. We found that modelling a strong continuous covariate as dichotomous is worse than dichotomizing a moderate covariate. The results also suggest that it is not a good idea to dichotomize continuous covariates in AUC regression adjustment.

These findings all lead to the same conclusion that there is an evidence that there is a significant chance that the DBP in the non-shock group is greater than the DBP in the shock group. The results were also consistent with those in the Monte Carlo simulations.

Finally, we conclude that incorrectly modeling covariates in AUC regression adjustment lead to unbiased estimates of the true effect.

CHAPTER 6: DISCUSSION

6.1 Conclusion

In the first part of this dissertation research, the primary objective was to evaluate the performance of propensity score methods to estimate the area under the ROC curve while controlling for confounding. The simulation study demonstrated that when AUC is used as measure of risk factor effect, conditioning on the propensity results in biased estimation of the true effect. When the true effect was null i.e. AUC was 0.5, matching on the propensity score and covariate adjustment on the propensity score were associated with less bias compared to the method of stratifying on the propensity score. When the true effect was different from the null effect, the estimated AUC were all associated with large bias for all different methods.

In a simulation study conducted by Austin et al. (2007), they found that controlling for covariates using propensity score methods when estimating conditional odds ratio and conditional hazard ratio resulted in biased estimation of the true effect (P. C. Austin et al., 2007) . Thus our results are not totally unexpected. This study is the first to evaluate the performance of different propensity score methods for estimating area under the ROC curve i.e. $P\left(Y_{RF} > Y_{NRF}\right)$. Due to the increased interest in epidemiologic research to report $P\left(Y_{RF} > Y_{NRF}\right)$ as the measure of association and to the use of propensity score methods to control for confounding, it is of practical importance that the statistical properties of propensity scores estimators for AUC be understood.

A secondary objective was to determine the best choice of variables to include in the propensity score model. We found that when matching and covariate adjustment on the propensity score methods are used, the propensity score model including variables associated

with outcome seems to have the least bias. Models including those variables that are both associated with outcome and risk group (these are referred to as true confounders) did not perform well. But these findings are not conclusive because the results were not consistent throughout the true effects and the amount of bias is still high. In prior research investigating the issue of variables selection in propensity score models, Brookhart et al. (2006) found that a propensity score model with only covariates associated with outcome or the true confounders resulted in a larger number of matched and a smaller mean squared error (Brookhart et al., 2006) Furthermore, Austin found that variables associated with treatment exposure but not the outcome increased the MSE of the estimated relative risk (Austin, 2008) .

A third objective was to compare the performance of the propensity score approach with that of a conventional regression approach to estimate $P\left(Y_{RF} > Y_{NRF}\right)$. The results of our simulation study show that the AUC regression model including all covariates associated with outcomes has the best performance and resulted in unbiased estimates of the risk effect. However, regression models that did not include all variables associated with outcome and only contained variables associated with risk factor group or variables associated with both risk group and outcome resulted in biased estimates of the true AUC and in an increased MSE. Austin et al. (2007) advocate that the choice between propensity score methods and regression adjustment when estimating odds ratio or hazard ratio should be based on whether one wishes to estimate the marginal or the conditional treatment effect. They noted that the conventional regression adjustment estimates conditional treatment effect while the propensity score estimates marginal treatment effects (P. C. Austin et al., 2007).

Finally, in the second part of this research, the goal was to investigate the impact of model misspecification in the conventional AUC regression adjustment. These modelling errors

include omitting covariates, dichotomizing continuous variables, modelling quadratic covariates as linear, and excluding interactions terms from the model. We found that the greatest bias was associated with the model that omitted a covariate strongly associated with the outcome. And in general, it is far more damaging to incorrectly model a strong covariate than to incorrectly model a covariate weakly associated with the outcome. The results of this study suggest that researchers must focus on these variables known to be strongly related to the outcome variable and should attempt to correctly model them.

Given these findings, we do not recommend the use of propensity score methods to provide adjusted estimates of $P\left(Y_{RF} > Y_{NRF}\right)$. Instead the conventional AUC regression adjustment is the method to use. When the outcome variable is continuous, if one is interested in using the propensity score methods, then the difference in means allows for unbiased estimation of the risk effect. When the conventional AUC regression adjustment is used to control for confounding, analysts must focus on variables related to the outcome; these covariates (especially if they are strongly associated with the outcome) should be correctly modelled in order to estimates accurate effect when assessing relationship between exposures and outcome. Furthermore, leaving out important variables in AUC regression models could lead to biased estimates of the true effect. Therefore, researchers and epidemiologists must make an effort to identify significant risk factors.

6.2 Limitations

A limitation to the use of the propensity score methodology in practice includes the fact that it only controls for observed variables. The unobserved variables are accounted for only if they are correlated with the observed covariates. Although the baselines covariates were

assumed to be correctly measured in the Monte Carlo simulations, this assumption in practice can be more problematic. Another limitation is in the choice of true AUC. The true effect of AUC was limited to 0.5, 0.7 and 0.9. Perhaps, the performance of the adjusted AUC should also be evaluated on a wider range of true effect. Other limitations include considering cases where the outcome variable is normally distributed. The effect of sample size on the performance of the proposed methods should have also been investigated. Here, only one sample size was used (N=500). Also, we assumed equal numbers of people with and without the risk factor. Different prevalences could have had an influence on results. Also, only independent variables were considered in this research; we could have considered correlated variables as well. We also looked at cases where the standard deviations between the risk and non-risk groups are equal; perhaps we should have also considered cases where the standard deviations between the two groups are not equal. Finally, all possible types of misspecification of the covariates have not been considered. Perhaps, other ways to model curvilinear association should have been considered.

6.3 Future Work

Future work in this area should focus on estimating the area under the curve under the non-normal assumption and on identifying if the propensity score methods performs well if we sought to estimate the "marginal" area under the ROC curve**.** This will involve defining a new measure based on the AUC used in the current research which will be referred to as the "marginal" area under the ROC curve (Austin 2007a). For future research, perhaps we should expand on nonlinear modelling errors beyond quadratic relationships; and on interaction models where the interactions terms are more independent i.e. not based on combination of strong, moderate and weak. In the future, we should also investigate the effect of sample size and the

prevalence of case-control on the performance of the adjusted AUC. For example, does it perform well for fairly small samples? We should also investigate the effect of prevalence of case and control in estimating AUC as well as the effect of different variance estimates between the two groups.

REFERENCES

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in medicine, 25*(4), 591-602. doi: 10.1002/sim.2256 [doi]

Afifi, A. A., & Azen, S. P. (1979). *Statistical analysis : a computer oriented approach* (2nd ed.. ed.). New York: New York : Academic Press.

Austin, P. C. (2007a). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in medicine, 26*(16), 3078-3094. doi: 10.1002/sim.2781 [doi]

Austin, P. C. (2007b). Propensity-score matching in the cardiovascular surgery literature from 2004 to 2006: a systematic review and suggestions for improvement. *The Journal of thoracic and cardiovascular surgery, 134*(5), 1128-1135. doi: S0022-5223(07)01243-3 [pii]

Austin, P. C. (2008). The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol, 61*(6), 537-545. doi: 10.1016/j.jclinepi.2007.07.011

Austin, P. C. (2009a). Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. *Biometrical journal.Biometrische Zeitschrift, 51*(1), 171-184. doi: 10.1002/bimj.200810488 [doi]

Austin, P. C. (2009b). Type I error rates, coverage of confidence intervals, and variance estimation in propensity-score matched analyses. *The international journal of biostatistics, 5*(1), Article 13-4679.1146. doi: 10.2202/1557-4679.1146 [doi]

Austin, P. C. (2009c). Using the Standardized Difference to Compare the Prevalence of a Binary Variable Between Two Groups in Observational Research. *Communications in Statistics*(Journal Article), 1228-1234.

Austin, P. C. (2010). The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Statistics in medicine, 29*(20), 2137-2148. doi: 10.1002/sim.3854 [doi]

Austin, P. C. (2013). The performance of different propensity score methods for estimating marginal hazard ratios. *Statistics in medicine, 32*(16), 2837-2849. doi: 10.1002/sim.5705 [doi]

Austin, P. C., Grootendorst, P., & Anderson, G. M. (2007). A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Statistics in medicine, 26*(4), 734-753. doi: 10.1002/sim.2580 [doi]

Austin, P. C., Grootendorst, P., Normand, S. L., & Anderson, G. M. (2007). Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med, 26*(4), 754-768. doi: 10.1002/sim.2618

Austin, P. C., & Mamdani, M. M. (2006). A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Statistics in medicine, 25*(12), 2084-2106. doi: 10.1002/sim.2328 [doi]

Begg, M. D., & Lagakos, S. (1990). On the consequences of model misspecification in logistic regression. *Environmental health perspectives, 87*(Journal Article), 69-75.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology, 163*(12), 1149-1156. doi: kwj149 [pii]

Brumback, L. C., Pepe, M. S., & Alonzo, T. A. (2006). Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in medicine, 25*(4), 575-590. doi: 10.1002/sim.2345 [doi]

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics, 24*(2), 295-313.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.. ed.). Hillsdale, N.J.: Hillsdale, N.J. : L. Erlbaum Associates.

D'Agostino, R. B., Jr. (1998). Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in medicine, 17*(19), 2265-2281. doi: 10.1002/(SICI)1097-0258(19981015)17:19<2265::AID-SIM918>3.0.CO;2-B [pii]

D'Agostino, R. B., Jr., & D'Agostino, R. B., Sr. (2007). Estimating treatment effects using observational data. *Jama, 297*(3), 314-316. doi: 297/3/314 [pii]

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics, 44*(3), 837-845.

Faraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *Journal of the Royal Statistical Society: Series D (The Statistician), 52*(2), 179-192. doi: 10.1111/1467-9884.00350

Greco, L., & Ventura, L. (2011). Robust inference for the stress–strength reliability. *Statistical Papers, 52*(4), 773-788. doi: 10.1007/s00362-009-0286-9

Green, D. M., & Swets, J. A. (1966). *Signal Detection theory and Psychophysics*. New York: Wiley & Sons, Inc.

Guttman, I., Johnson, R. A., Bhattacharyya, G. K., & Reiser, B. (1988). Confidence Limits for Stress-Strength Models With Explanatory Variables. *Technometrics, 30*(2), 161-168. doi: 10.1080/00401706.1988.10488363

Hauck, W. W., Hyslop, T., & Anderson, S. (2000). Generalized treatment effects for clinical trials. *Statistics in medicine, 19*(7), 887-899. doi: 10.1002/(SICI)1097-0258(20000415)19:7<887::AID-SIM388>3.0.CO;2-L [pii]

Hoeffding, W. (1948). A Non-Parametric Test of Independence. *The Annals of Mathematical Statistics, 19*(4), 546-557. doi: 10.1214/aoms/1177730150

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley & Sons, Inc.

Janes, H., Longton, G., & Pepe, M. (2009). Accommodating Covariates in ROC Analysis. *The Stata journal, 9*(1), 17-39.

Lagakos, S. W. (1988). Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. *Statistics in medicine, 7*(1-2), 257-274.

Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics, 18*(Journal Article), 50-60.

Merck. (2009). *The Merck manual home health handbook* (Vol. [3rd ed.]..). Whitehouse Station, NJ: Whitehouse Station, NJ : Merck Research Laboratories.

Monson, R. R. (1990). Occupational Epidemiology. (Journal Article).

Normand, S. T., Landrum, M. B., Guadagnoli, E., Ayanian, J. Z., Ryan, T. J., Cleary, P. D., & McNeil, B. J. (2001). Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *Journal of clinical epidemiology, 54*(4), 387-398. doi: S0895435600003218 [pii]

Nunney, I., Clark, A., & Shepstone, L. (2013). Estimating treatment effects in a two-arm parallel trial of a continuous outcome. *Statistics in medicine, 32*(6), 941-955. doi: 10.1002/sim.5556 [doi]

Osganian, S. K., Stampfer, M. J., Spiegelman, D., Rimm, E., Cutler, J. A., Feldman, H. A., . . . Nader, P. R. (1999). Distribution of and factors associated with serum homocysteine levels in children: Child and Adolescent Trial for Cardiovascular Health. *Jama, 281*(13), 1189-1196. doi: joc81454 [pii]

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics, 54*(1), 124-135.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

Pepe, M. S., & Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics, 60*(2), 528-535. doi: 10.1111/j.0006-341X.2004.00200.x [doi]

Pepe, M. S., & Longton, G. (2005). Standardizing diagnostic markers to evaluate and compare their performance. *Epidemiology (Cambridge, Mass.), 16*(5), 598-603. doi: 00001648-200509000-00002 [pii]

Perkins, S. M., Tu, W., Underhill, M. G., Zhou, X. H., & Murray, M. D. (2000). The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiology and drug safety, 9*(2), 93-101. doi: 10.1002/(SICI)1099-1557(200003/04)9:2<93::AID-PDS474>3.0.CO;2-I [doi]

Pett, M. A. (1997). *Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions.* Thousand Oaks, CA: Sage Publications.

Reiser, B., & Guttman, I. (1986). Statistical inference for Pr(Y<X): the normal case. *Technometrics, 28(3)*(Journal Article), 253-257.

Rosenbaum, P. R. (1995). *Observational studies*. New York: New York : Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 7*(Journal Article), 41-55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(Journal Article), 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39*(Journal Article), 33-38.

Rosenthal, J. A. (1996). Qualitative Descriptors of Strength of Association and Effect Size. *Journal of Social Service Research, 21*(4), 37-59. doi: 10.1300/J079v21n04_02

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: relating theory to practice. *Biometrics, 52*(1), 249-264.

Schroeder, H. A. (1966). Municipal drinking water and cardiovascular death rates. *Jama, 195*(2), 81-85.

Setoguchi, S., Schneeweiss, S., Brookhart, M. A., Glynn, R. J., & Cook, E. F. (2008). Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety, 17*(6), 546-555. doi: 10.1002/pds.1555 [doi]

Stone, R. A., Obrosky, D. S., Singer, D. E., Kapoor, W. N., & Fine, M. J. (1995). Propensity score adjustment for pretreatment differences between hospitalized and ambulatory patients with community-acquired pneumonia. Pneumonia Patient Outcomes Research Team (PORT) Investigators. *Medical care, 33*(4 Suppl), AS56-66.

Tian, L. (2008). Confidence intervals for P(Y1>Y2) with normal outcomes in linear models. *Statistics in medicine, 27*(21), 4221-4237. doi: 10.1002/sim.3290 [doi]

Walsh, S. J. (1997). Limitations to the robustness of binormal ROC curves: effects of model misspecification and location of decision thresholds on bias, precision, size and power. *Statistics in medicine, 16*(6), 669-679. doi: 10.1002/(SICI)1097-0258(19970330)16:6<669::AID-SIM489>3.0.CO;2-Q [pii]

Weitzen, S., Lapane, K. L., Toledano, A. Y., Hume, A. L., & Mor, V. (2004). Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiology and drug safety, 13*(12), 841-853. doi: 10.1002/pds.969 [doi]

Westreich, D., Cole, S. R., Funk, M. J., Brookhart, M. A., & Sturmer, T. (2011). The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiology and drug safety, 20*(3), 317-320. doi: 10.1002/pds.2074 [doi]

Woo, M., Reite, J. P., & Karr, A. F. (2008). Estimation of propensity scores using generalized additive models. *Statistics in.medicine*(Journal Article).

Wooldridge, J. M. (2000). *Introductory econometrics : a modern approach*. Cincinnati, OH: Cincinnati, OH : South-Western College.

APPENDIX

A. BALANCE DIAGNOSTICS & SIMULATION CHECK

The correctness of the simulated data was checked by, for a single dataset of size N=500, whether:

a. Approximately 50% of subjects are exposed to the risk factor by computing the frequency of the risk factor status.

Risk Factor Status		
T	Frequency	Percent
0	251	50.2
1	249	49.8

b. The covariates are imbalanced at baseline by computing a standardized difference between subjects with risk factor and subjects without risk factor for each covariate in the data.

Table A.1 Standardized difference comparing the mean or prevalence of baseline covariates between risk factor groups

Continuous covariates	**Standardized difference**	**Binary covariates**	**Standardized difference**
c1	0.337	b1	0.286
c2	0.182	b2	0.169
c3	0.001	b3	0.004
c4	0.331	b4	0.288
c5	0.182	b5	0.166
c6	0.000	b6	0.002
c7	0.331	b7	0.287
c8	0.180	b8	0.165
c9	0.001	b9	0.002

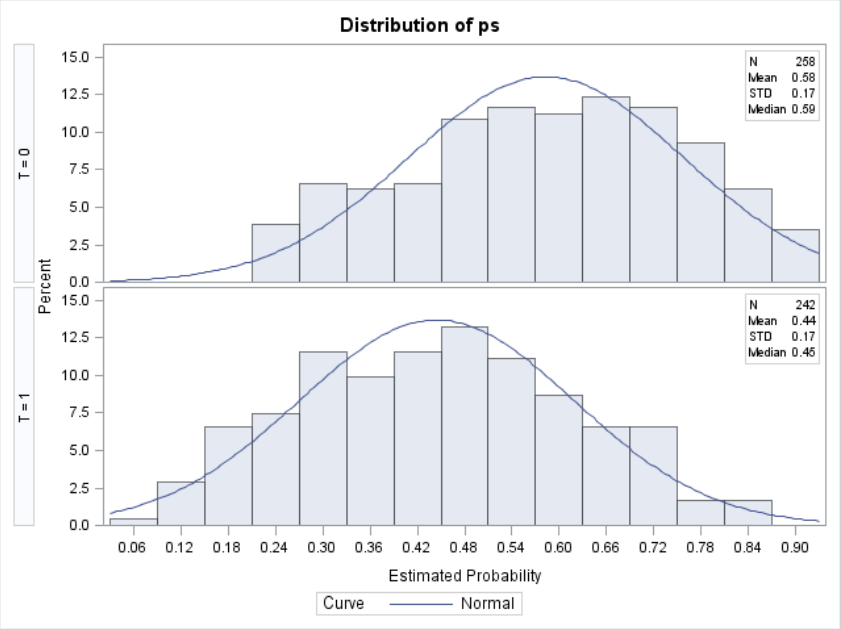
A standardized difference greater than 0.1 is considered as a significant difference in the mean or prevalence of a covariate between risk factor groups (Normand et al., 2001).

c. The distribution of the propensity score "reasonably" overlap by **c**omputing a c-statistic to predict if the distributions of the PS overlap (Westreich, Cole, Funk, Brookhart, & Sturmer, 2011). Westreich reported a high-c statistic in the propensity model is "neither necessary nor sufficient for control of confounding".

The values of the C-statistic in the propensity score model including variables related to treatment; all covariates, binary covariates only, and continuous covariates only were 0.753, 0.756, 0.666 and 0.696, respectively. These values of the c-statistic are considered "reasonable" since they are neither too high nor too low. (Recall the c-statistic takes on values between 0.5 and 1).

d. The two risk factor groups are comparable i.e. if the overall distribution of the estimated propensity score within each risk group "reasonably" overlap.. We checked this via histograms.

Figure A.1 Distribution of the estimated propensity score in each risk factor group



The estimated propensity score "reasonably" overlap which means the risk factors and the non-risk factors groups are comparable.

e. Balance for the measured covariates is achieved between the risk factor and the non-risk factor groups by 1) Assessing the balance of each covariate after adjustment by computing a standardized difference; and 2) Summarizing the distribution of the propensity scores via box plots: If they overlap then a good balance is achieved.

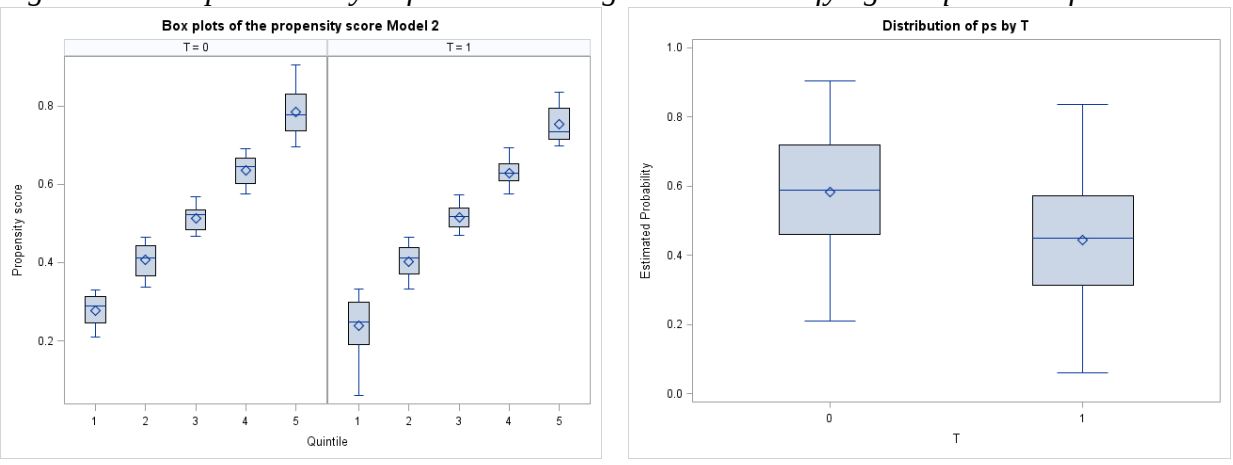
This property is checked using two propensity score methods: 1) Propensity score matching to check balance of the covariates after adjustment in the PS matched sample and 2) Stratification on the propensity using the technique of box plots.

Table A.2 Standardized difference comparing the mean or prevalence of variables between risk factor groups after PS adjustment by matching technique

Continuous covariates	Standardized difference	Binary covariates	Standardized difference
c1	0.012	b1	0.039
c2	0.095	b2	0.013
c3	0.053	b3	0.105
c4	0.009	b4	0.039
c5	0.066	b5	0.026
c6	0.089	b6	0.184
c7	0.058	b7	0.013
c8	0.035	b8	0.026
c9	0.046	b9	0.079

After adjustment in the propensity score matched sample, all the covariates have a standardized difference less than 0.1 except b6. Hence, balance is achieved for almost 95% of the baseline variables.

Figure A.2 Graphical analysis for balance diagnostics in stratifying the quintiles of the PS



The distribution of the estimated propensity score appears to be similar within the two risk groups. Given this graphical evidence, we conclude that stratifying on the quintiles of the propensity scores resulted in the creation of subjects who are balanced in observed covariates between the two risk factor groups.

f. The association between the risk factor group and the covariates depicts an odds ratio of log(2) and log(1.5) for binary variables and log(1.5) and log(1.25) for continuous variables by computing the odds ratio between the risk factor and each covariate in the data.

Table A.3 Odds Ratio Estimates of the simulated data

Odds Ratio Estimates			
Variable	**Point estimate**	**Variable**	**Point estimate**
b1	2.00	c1	1.51
b2	1.51	c2	1.25
b3	1.01	c3	1.00
b4	2.01	c4	1.50
b5	1.51	c5	1.25
b6	1.00	c6	1.00
b7	2.00	c7	1.50
b8	1.50	c8	1.25
b9	1.00	c9	1.00

We clearly see that $(b_1, b_4, b_7) = 2$, $(b_2, b_5, b_8) = 1.5$, $(c_1, c_4, c_7) = 1.5$, and $(c_2, c_5, c_8) = 1.25$. This is what was expected. Hence, our data have been correctly generated.

g. The association between the outcome and the covariates depicts a correlation of 0.5 for strong and 0.3 for moderate by computing the correlations between the independent variables and the outcome.

Table A.4 Correlation coefficients between outcome and the simulated continuous covariates

	Outcome Y								
c1	0.498	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001	-0.001
c2	0	0.498	0	0	-0.001	0	0	0	0
c3	-0.001	-0.001	0.498	-0.001	-0.002	-0.002	-0.001	-0.001	-0.001
c4	0	-0.001	-0.001	0.309	-0.001	-0.001	-0.001	-0.001	-0.001
c5	0.001	0	-0.001	0	0.309	0	0	0	0
c6	0.001	0	-0.001	0	0	0.309	0	0	0
c7	0	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
c8	0	0	0	0	0	0	0	0	0
c9	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001	0.001

Table A.5 Correlation coefficients between outcome and the simulated dichotomous covariates

	Outcome Y								
b1	0.497	-0.001	-0.003	-0.002	-0.002	-0.002	-0.002	-0.002	-0.002
b2	0.001	0.499	0	0	0	0	0	0	0
b3	0	0.001	0.499	0.001	0.001	0.001	0.001	0.001	0.001
b4	0	0	0	0.309	0	-0.001	0	0	0
b5	-0.001	0	0	0	0.309	0	0	0	0
b6	0	0	0	0	0	0.309	0	0	0
b7	0	0	0	0	0	0	0	0	0
b8	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
b9	0	0	0	0	0.001	0	0	0	0

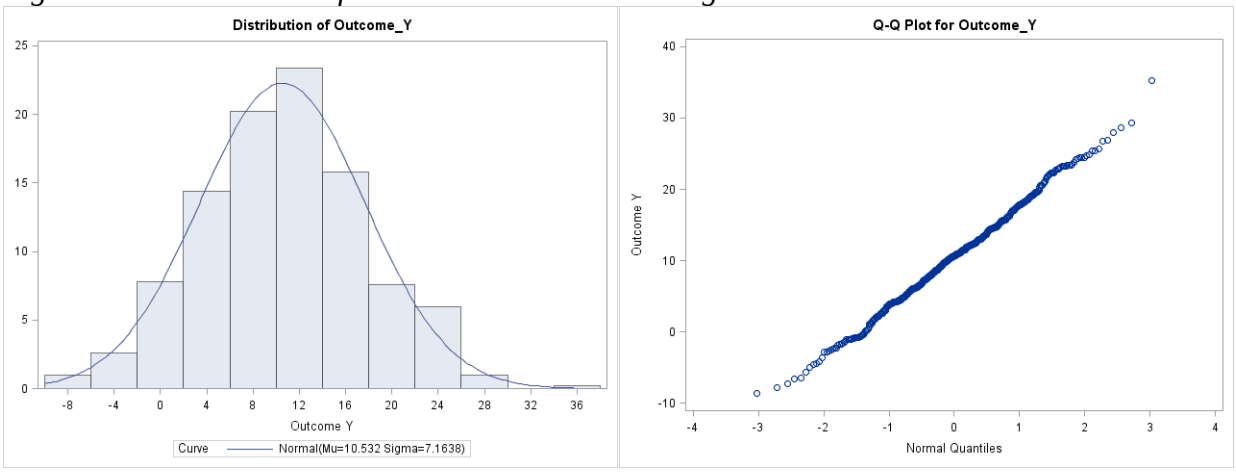
Table A.6 Correlation coefficients between the simulated covariates

	b1	b2	b3	b4	b5	b6	b7	b8	b9
c1	0	0.001	0	0.001	-0.001	0.001	0.001	0	0.001
c2	0	-0.001	-0.002	0	0	-0.002	0	-0.002	0.001
c3	-0.001	0.001	0	0.001	0.001	0	0.002	0.001	0
c4	-0.001	0	0.001	0	0.001	-0.002	0	0	-0.001
c5	0.001	-0.002	0.001	-0.001	0	0.004	-0.001	0	-0.001
c6	-0.001	0.001	0	0.001	0	0	-0.001	0	0
c7	0	-0.001	-0.001	0	0	0.001	0.001	-0.001	0.002
c8	0.001	0	-0.001	-0.001	0	-0.001	-0.001	0	0
c9	0	0.001	0	0.001	-0.002	0	-0.002	0	0

These correlations values are expected. Furthermore, the correlations values in Table 3-10 show that the covariates are uncorrelated with each other thus independent. This demonstrates once again that our data has been correctly generated.

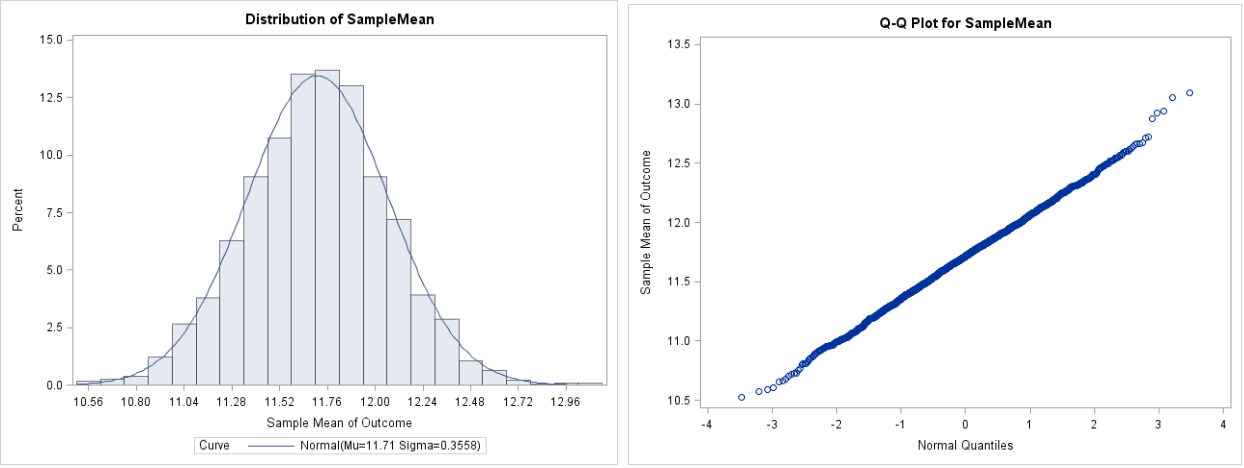
h. The outcome based on a single dataset is approximately normal by 1) Overlaying a normal PDF on a histogram; and 2) Constructing a Q-Q plot. If the data are sample from the normal distribution, then the points on the plot tend to fall along a straight line (Chambers et al. 1983).

Figure A.3 Distribution of the outcome based on a single dataset



i. The sample mean AUC based on 2500 replications is approximately normal by 1) Overlaying a normal PDF on a histogram; and 2) Constructing a Q-Q plot. If the data are sample from the normal distribution, then the points on the plot tend to fall along a straight line (Chambers et al. 1983).

Figure A.4 Distribution of the AUC sample mean across 2500 datasets



B. SAS CODES

B.1 SAS Code to compute the "Stratified "adjusted AUC

```
*Using Stratification on the PS, for each PS model, we will compute the stratum specific AUC
along with its std. error;

%macro StratifiedAUC (data,title,outcome);
proc sort data =&data; by quintile; run;
proc logistic data = &data;
        ods select none ;
        by quintile;
        model anyshock  = &outcome; *(event='Women');
        roc 'DBP2' &outcome; *To output auc and its std error;
        ods output ROCAssociation= Raucs;
        ods output ResponseProfile= Rfreq;
        Title "Logistic model to estimate the stratum specific AUC for: &title";
Run;

Data AUCStratum (Keep = Quintile Area StdErr Outcome Count);
        set Raucs;
        set Rfreq;
        call symput ('AUC', area);
        call symput ('StdErr', StdErr);
        call symput ('Shock', Count);
        call symput ('Noshock', Count);
run;

Data AUCStratum ;
        set AUCStratum;
        length NewShock $14;
        if Outcome = 0 then NewShock = 'Shock'; else NewShock = 'Noshock';
run;

*The adjusted AUC and its std error is caluclated using the weighted average of the stratum
specific AUCs: Ws = ms*ns/sum(ms*ns);
proc sql;
        title "Adjusted AUC as the weighted average of the stratum-specific AUCs";
        create table Abc as
                select one.Quintile,
                        one.Count as shock, two.Count as Noshock, one.area as auc, one.StdErr as
stderr,
                        one.Count*two.Count as  numerator,
                        sum(one.Count*two.Count) as denominator,
                        one.Count*two.Count/sum(one.Count*two.Count)as weight,
```

```
                    one.area*(one.Count*two.Count/sum(one.Count*two.Count)) as
StratumAUC
            from AUCStratum as one INNER JOIN AUCStratum as two
            On (one.Quintile = two.Quintile)
            where one.NewShock = 'Shock' and two.NewShock = 'Noshock';
quit;

proc sql;
    create table Aa as
        select *,
            sum(StratumAUC) as AdjAUC,
            weight*stderr as StratumSE,
            sum(calculated StratumSE) as AUCSE,
            (calculated AdjAUC) - 1.96*(calculated AUCSE)as LLCI,
            (calculated AdjAUC) + 1.96*(calculated AUCSE)as ULCI
        from Abc;
Quit;

Data StratifiedAUC (keep = AdjAUC AUCSE LLCI ULCI);
    set AA;
    where Quintile = 1;
ods select all;
proc print data = StratifiedAUC noobs; format AdjAUC 5.4 AUCSE 5.4 LLCI 5.4 ULCI 5.4 ;
    title "The Adjusted Stratified AUC is for: &title";
run;
%mend;

* Example of macro call;
%StratifiedAUC(Ps_dataM1, PS model 1, dbp2);
```

B.2 SAS Code to compute the adjusted AUC using the concept of placement values

```
/*
The aAUC %macro is based the COMPROC command developed by
Janes, Longton & Pepe, 2008 for accommodating covariates in ROC analysis.

Description
The aAUC macro estimates the area under the ROC curve using the concept of placement values
while adjusting for covariates.
The placement values (PV) are estimated parametrically assuming a normal distribution. The
process is conducted in two steps. First, estimate the cumulative distribution (CDF) of the
response Y in the control group as a function of Z (i.e. the covariates of interest to adjust for).
This is done by specifying a linear model (Y = Bo + B1Z + e) assuming the error term is
normally distributed and the covariates act linearly on the distribution of Y. Then for each
subject i in the risk factor (or “disease”, or “treatment”, or “case”, or “condition” or “event”)
group, compute the placement values. The PV is the standard normal CDF of (Y - Bo_hat -
```

B1_hatZ)/sd, where Bo_hat, B1_hat, and sd are the regression coefficients estimates and the standard deviation of the control observations, respectively. The second step is to estimate the adjusted AUC by computing the mean of the estimated placement values.

The inputs are: data, T, outcome adjcov, bsamp, n

a- data = specifies the dataset to be used for analysis.
b- T = specifies the variable denoting the risk factor (or "case" or "disease" or "treatment") group. T has the values 0/1.
c- outcome = specifies the continuous response arising from the populations with and without the risk factor.
d- adjcov = specifies the covariates to adjust for.
e- n = number of covariates to adjust for.
f- bsamp = number of bootstrap samples to be drawn for estimating standard errors and Cis of the the adjusted AUC.

*/

%macro aAUC (data, T, outcome, adjcov, n, bsamp);

```
*
+*************************************************************************+
 Computing the placement values and the adjusted AUCs for each replicate (SampleID)
*
+*************************************************************************+;

* +----------------------------------------------------------------------+
 *Step 1: running an OLS regression to estimate the CDF of Y in the control group
* +----------------------------------------------------------------------+;

proc reg data=&data (where =(&T =0))outest=OutEst;
        By SampleID;
        model &outcome = &adjcov/noprint;
        title "Linear model to estimate the CDF of Y in the control group";
run;

*Read in the data where only the regression coefficients are computed and "renaming" the
variables;
Data RegCoef (Drop = _MODEL_ _DEPVAR_ &outcome &adjcov i) ;
        set OutEst;
        if _TYPE_ = 'PARMS';
        array cov{&n} &adjcov;
        array coef{&n} coef1-coef&n;
                do i = 1 to &n;
                        coef{i}=cov{i};
                end;
run;
```

```
*Preparing the case data and creating a new variable _Type_ to merge on;
Data Case (Keep = SampleID &T &outcome _TYPE_ &adjcov);
        Set &data (where =(&T =1));
        _TYPE_ = 'PARMS';
run;

*Merging the regression coefficients data with the case data;
data CaseData ;
        merge RegCoef Case;
        by _TYPE_;
run;

* +-----------------------------------------------------------------------+
  Step 2: Estimating the placement values for each subject in the risk factor.
* +-----------------------------------------------------------------------+;

data PVdata (Keep = SampleID &T &outcome _RMSE_ Intercept z1-z&n meancontrol PV);
        set CaseData;
        array cov{&n} &adjcov;
        array coef{&n} coef1-coef&n;
        array covcoeff{&n} z1-z&n;
                do i =1 to &n;
                        covcoeff{i} = cov{i}*coef{i};
                end;
        meancontrol = intercept + sum(of covcoeff {*});
        PV = PROBNORM((&outcome - meancontrol)/_RMSE_);
run;

* +-----------------------------------------------------------------------+
  Step 3: Estimating the mean of the placement values i.e. the Adjusted AUC.
* +-----------------------------------------------------------------------+;

proc means  data = PVdata noprint;
        By sampleID;
        var PV;
        output out =  AdjAUC0 mean= AdjAUC var = AUCVar std = AUCStd;
run;

* +*************************************************************************+
   Computing the variance and standard deviation for each adjusted AUC via Bootstrap
*
+*************************************************************************+;
proc sort data = &data;
        by SampleID T;
run;
```

```
%let MyData = &data;
proc surveyselect data=&MyData seed=1
        out=BootSS
   method=urs samprate=1
   reps=&bsamp
   outhits;
        strata sampleID T; *To maintain prevalence of case and control as in the original data;
run;

* Redoing steps 1, 2, 3 above with the bootstrap sample;

* +-----------------------------------------------------------------------+
 *Step 1: running an OLS regression to estimate the CDF of Y in the control group
* +-----------------------------------------------------------------------+;

proc reg data=BootSS (where =(&T =0))outest=OutEst1;
        By SampleID;
        model &outcome = &adjcov/noprint;
        title "Linear model to estimate the CDF of Y in the control group";
run;

*Preparing the case data and creating a new variable _Type_ to merge on;
Data Case1 (Keep = Replicate SampleID &T &outcome _TYPE_ &adjcov);
        Set BootSS (where =(&T =1));
        _TYPE_ = 'PARMS';
run;

*Read in the data where only the regression coefficients are computed and "renaming" the
variables;
Data RegCoef1 (Drop = _MODEL_ _DEPVAR_ &outcome &adjcov i) ;
        set OutEst1;
        if _TYPE_ = 'PARMS';
        array cov{&n} &adjcov;
        array coef{&n} coef1-coef&n;
                do i = 1 to &n;
                        coef{i}=cov{i};
                end;
run;

*Merging the regression coefficients data with the case data;
data CaseData1 ;
        merge RegCoef1 Case1;
        by _TYPE_;
run;
```

```
* +-----------------------------------------------------------------------+
  Step 2: Estimating the placement values for each subject in the risk factor.
* +-----------------------------------------------------------------------+;

data PVdata1 (Keep = Replicate SampleID &T &outcome _RMSE_ Intercept z1-z&n
meancontrol PV);
        set CaseData1;
        array cov{&n} &adjcov;
        array coef{&n} coef1-coef&n;
        array covcoeff{&n} z1-z&n;
                do i =1 to &n;
                        covcoeff{i} = cov{i}*coef{i};
                end;
        meancontrol = intercept + sum(of covcoeff {*}s);
        PV = PROBNORM((&outcome - meancontrol)/_RMSE_);
run;

proc datasets library=work;
        delete CASE1 CASEDATA1;
run;
quit;
* +-----------------------------------------------------------------------+
  Step 3: Estimating the mean of the placement values i.e. the Adjusted AUC.
* +-----------------------------------------------------------------------+;
proc sort data = PVdata1; by sampleID Replicate ; run;
proc means  data = PVdata1 noprint;
        By sampleID Replicate ;
        var PV;
        output out =  AdjAUCint mean= AdjAUC var = AUCVar std = AUCStd;
run;
proc means  data = AdjAUCint VARDEF = N noprint;
        By sampleID ;
        var AdjAUC;
        output out =  AdjAUC1 mean= AdjBootMean var = AUCBootVar std = AUCBootStd;
run;
Data adjustedAUC;
        merge adjauc0 adjauc1;
        By SampleID;
        T_crit = tinv(1-&alphalev/2, &bsamp-1);
        LB_N = AdjAUC - T_crit*AUCBootStd; *Normal Distribution CI;
        UB_N = AdjAUC + T_crit*AUCBootStd;
run;
%mend aAUC;

* Example of macro call;
%aAUC(shockdata, shockStatus, dbp2, SBP MAP HR DBP CI MCT UO, 7, 1000);
```

VITAE

Hadiza Galadima was born and raised in Niamey, in the Republic of Niger. She holds a D.E.U.G (Diplôme d'Etudes Universitaires Générales) degree in Physics and Chemistry from Université Hassan II Mohammedia, in Morocco. She moved to the US in 2001 where she first completed an Associate Degree in Graphic Design at Washtenaw Community College; then a bachelor degree in Statistics and Actuarial Science where she graduated Magna Cum Laude from St. Cloud State University, in Minnesota. She started a Master's degree in Applied Statistics at St. Cloud State University but after a year in the program she decided to pursue with a PhD in Biostatistics.

She joined the department of Biostatistics in the Medical School of Virginia Commonwealth University (VCU) in fall 2010. Throughout her time at VCU, She has held numerous positions such as teaching assistant, intern at UNOS and a research assistant in the department of Healthcare Policy and Research. Her dissertation focuses on examining the performance of methods to control for confounding when association is quantified by area under the ROC curve. Her research interests and expertise include observational studies, propensity score analysis, AUC, variables selection for propensity score model building, model misspecification and disparities research as it relates to HIV/AIDS.

On her spare time, she likes working on various graphic design projects ranging from making logos, flyers & business cards, to taking photos and creating web pages.