# USE OF AREA UNDER THE CURVE (AUC) FROM PROPENSITY MODEL TO ESTIMATE ACCURACY OF THE ESTIMATED EFFECT OF EXPOSURE

by

Zhijiang Zhang

B. Med, Nanjing Medical University, China, 1998

M.Sc, Harvard University, 2003

Submitted to the Graduate Faculty of

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2007

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

**Zhijiang Zhang**

It was defended on

**July 11, 2007**

and approved by

Thesis Advisor:
**Andriy Bandos, PhD**
Research Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Committee Member:
**Kevin E. Kip, PhD**
Associate Professor
Department of Epidemiology
Graduate School of Public Health
University of Pittsburgh

Committee Member:
**Gong Tang, PhD**
Assistant Professor
Department of Biostatistics
Graduate School of Public Health
University of Pittsburgh

Andriy Bandos, PhD

# USE OF AREA UNDER THE CURVE (AUC) FROM PROPENSITY MODEL TO ESTIMATE ACCURACY OF THE ESTIMATED EFFECT OF EXPOSURE

Zhijiang Zhang, M.S.

University of Pittsburgh, 2007

**Objective:** To investigate the relationship between the area under the Receiver Operating Characteristic curve (AUC) of the propensity model for exposure and the accuracy of the estimated effect of the exposure on the outcome of interest.

**Methods:** A Monte Carlo simulation study was performed where multiple realizations of three binary variables: outcome, exposure of interest and a covariate were repeatedly generated from the distribution determined by the parameters of the "propensity" and "main" models and the prevalence of the exposure. "Propensity" model was a logistic regression with the exposure of interest as a dependent variable and a single covariate as an "independent" variable. "Main" model was a logistic regression with outcome as a dependent variable, exposure of interest and covariate as "independent" variables. A total of 500 simulations were performed for each considered combination of the model parameters and the prevalence of the exposure. AUC was estimated from the probabilities predicted by the propensity score model. The accuracy of the estimated effect of exposure was primarily assessed with the square root of Mean Square Error (RMSE); the fifth and ninety-fifth percentile of the empirical distribution of the estimator were used to illustrate a range of not unlikely deviations from the true value.

**Results:** The square root of Mean Square Error of the estimated effect of exposure increases as AUC increases from 0.6 to 0.9. Varying values for parameters of the propensity score model or the main effect model does not change the direction of this trend. As the proportion of exposed subjects changes away from 0.5 the RMSE increases, but the effect of AUC on RMSE remains approximately the same. Similarly, as sample size changes from 50 to 100 or 200, the RMSE of effect estimate decreases on average, but the effect of AUC on RMSE remains approximately the same. Also, the rate of change in RMSE increases with increasing AUC; the rate is the lowest when AUC changes from 0.6 to 0.7 and is highest when AUC changes from 0.8 to 0.9.

**Conclusions:** The AUC of the propensity score model for exposure provides a single, relatively easy to compute, and suitable for various kind of data statistic, which can be used as an important indicator of the accuracy of the estimated effect of exposure on the outcome of interest. The public health importance is that it can be considered as an alternative to the previously suggested (Rubin, 2001) simultaneous consideration of the conditions of closeness of means and variances of the propensity scores in the different exposure groups. Our simulations indicate that the estimated effect of exposure is highly unreliable if AUC of the propensity model is larger than 0.8; at the same time AUCs of less than 0.7 are not associated with any substantial increase of inaccuracy of the estimated effect of exposure.

## TABLE OF CONTENTS

# LIST OF TABLES

## LIST OF FIGURES

# 1.0 INTRODUCTION

In observational studies, other risk factors may be correlated with the exposure of interest due to lack of randomization [1, 2]. Multivariable modeling attempts to solve this problem by putting the exposure of interest together with other measured covariates in the model. The adjustment works well under many circumstances. However, when the exposure of interest is highly correlated with other covariates, the estimated effect of exposure on outcome may become inaccurate [3, 4]. A strong association between the exposure of interest and other covariates can reveal itself through different phenomena such as collinearity, quasicomplete separation, or zero cells.

While there are specific techniques to flag the problems of collinearity, quasicomplete separation and zero cells, one can try to identify directly the underlying problem of a strong association. Perhaps one of the best approaches to determine the degree of the association between the exposure and other covariates in the collected data is to fit a regression model with the exposure as dependent variable and other covariates as independent variables. Under such an approach the degree of the association between the exposure of interest and other covariates is directly related to the ability of the model to "predict" the exposure based on the values of the other covariates.

A model of the probability of exposure (or specific treatment assignment) via other covariates is often used in observational studies and it typically terms as "propensity model".

Usually logistic regression is used as a propensity model, and we will term the probabilities predicted by the propensity model as estimated propensity scores. The distribution of the propensity scores determines the ability of the propensity model to "predict" the exposure of interest, and hence characterizes the degree of the association between the exposure of interest and other covariates. Since the degree of such an association is closely related to the trustworthiness (accuracy) of the estimated relationship between the exposure and the outcome of interest, certain criteria have been suggested to assess the adequacy of the estimated effect of the exposure based on the estimated propensity scores (Rubin, 2001). However, these criteria imply normal distributions. On the other hand area under the curve (AUC) is an index that reflects the discrimination ability of the logistic model regardless of the distribution of the predicted probabilities. In this work we investigate how well the AUC of the propensity model reflects the trustworthiness of effect estimator for the exposure of interest in the main effect model.

## 1.1 COLLINEARITY AND QUASICOMPLETE SEPARATION

One of the components of the association of other covariates with the exposure of interest is known as collinearity or near-collinearity. In general, near-collinearity occurs when two or more confounders and/or exposure are highly correlated with each other, sometimes even to the extent when it is difficult or even impossible to distinguish their individual influence on the outcome [5]. Collinearity can be defined for any two sets of continuous covariates. But here we are more interested in the situation when near collinearity exists between the exposure of interest and other covariates in an observational study. There are several approaches to flag near-collinearity at the analysis stage. When near-collinearity exists, the regression coefficients are likely to change dramatically in either magnitude or sign according to whether the other confounders are included or not; the standard errors of the regression coefficients may become substantially higher, or nonsignificant despite a high $R^2$. In logistic regression, one can look for estimated coefficients of unreasonable magnitude or with estimated standard error which is much larger than expected [6]. Ridge logistic regression or principal components logistic regression are alternatives to standard logistic regression when near-collinearity exists [3, 4]. The simplest way to check collinearity between continuous covariates is pairwise correlation analysis. However, this method does not necessarily detect multicollinearity. A better approach is to fit regression model with each covariate as dependent variable and the other variable as independent variable. In practice VIF (variance inflation factor) is also used, which is defined as $(1- R^2)^{-1}$. It is suggested that largest VIF>10 or mean VIF>1 indicate collinearity problem. In present study, we are primarily interested in the collinearity between covariates and binary exposure on accuracy of effect measure for exposure.

Another problem related to the association of the exposure with the other covariates of interest is “quasicomplete separation”, or “not adequately overlapping covariates”. This problem happens when a third covariate, other than exposure of interest and outcome, have a distribution across the exposure/treatment groups without or with little overlapping [6]. For continuous covariates, it implies that the values are greatly different in one group than in the other group. For example, subjects in one group can be mostly below 20 years of age, while in the other group they are mostly over 35 years of age. For categorical covariates, one exposure group can include subjects of almost exclusively one category, such as male; while the other exposure group includes subjects almost exclusively from the other categories, such as female. Naturally, this problem of “not adequately overlapping” is more likely to happen when the sample size is small, the proportion of exposure or treatment is far from 0.5, or the number of covariates is large [6].

Near-collinearity, quasi-complete separation and zero cells are interconnected but have distinct features when considered in the relationship to the two sets of multi-categorical (e.g. continuous) variables. In application to the binary exposure level and a set of covariates, the distinctions fade away and it becomes easier to refer to the corresponding problem collectively as a degree of association between level of the exposure and other covariates.

In the presence of a strong association between the exposure of interest and other covariates it is difficult to compare the exposure groups. If the effects of the covariates are ignored, the estimated effect of exposure might be highly inaccurate due to incorporation of the effects of unaccounted covariates. For example, when there are much more old people in one exposure group than in the other exposure group, the difference of the outcome between exposure groups can be attributed to either exposure or age or both. Or, when one exposure group is consisted of substantially more male subjects, while the other exposure group is consisted of substantially more female subjects, the difference of the outcome between exposure groups can be either the result of different exposure or the result of gender effect. Unfortunately, in the presence of mentioned near-collinearity analytical adjustments do not eliminate the problem completely. Because of the binary nature of the exposure considered in this woek, both problem can be classified simultaneously as near-collinearity and quasi-complete separation.

There are several different approaches to diagnosing the problems of strong association, such as large estimates in the target model, or severe imbalance in a univariate analyses. Propensity score modeling can be viewed as a mechanism to model such association directly, and thus, provides yet another, and perhaps more natural, approach for diagnosing potential problems related to strong association.

## 1.2 PROPENSITY SCORES

Propensity score modeling was proposed by Rosenbaum and Rubin in 1983 [7]. The propensity score is the conditional probability of “being assigned” to a particular exposure, given a set of observed characteristics. It can be estimated from a logistic regression, with the exposure as the dependent variable and the potential confounders as “independent” variables [8]. Patients with the same propensity score have equal estimated probabilities to “be assigned” to each exposure group and the same conditional distribution of the observed characteristics [7, 9, 10]. Therefore, it is akin to randomized clinical trials, which achieve balance of confounders between the exposure groups through the process of randomization [11]. However, propensity score can not control for unknown or unobserved confounders, whereas randomization can stochastically balance both observed and unobserved confounders [11, 12]. Some researchers have suggested methods to evaluate sensitivity of the propensity model to unknown confounders [13].

The technique using propensity score for adjustments can be classified into three types: matching, subclassfication, and weighting. Matching is the paring of exposed units and unexposed units with similar values of the propensity score. All unmatched units will be discarded. One-one Mahalanobis metric matching within propensity score calipers is the most popular method, but one treated unit matching multiple unexposed units is also proposed [12]. Subclassification creates subclasses of exposed or unexposed units with similar values of propensity score. First rank all units by their propensity score values and then use boundaries to create five or six subclasses, within which there are approximately the same total number of units [12]. Weighting methods weight each exposed unit with the inverse of propensity score, and weight each unexposed unit with the inverse of one minus the propensity score [12]. All these methods do not involve outcome variables, so these efforts will not affect the effect estimator on outcome, analogous to the way randomization works for clinical trial. Matching results in well-balanced but smaller groups for comparison. Subclassification retains a larger sample size, but the exposure groups are more heterogeneous within each subclass. Another application of propensity score is the use as a covariate for adjustment in a multivariate regression model, with or without inclusion of other potential confounders.

In addition to using propensity scores for the adjustment there is another very important utility of the propensity scores, specifically on the initial stages of analysis they can be used to diagnose if successful balance has been achieved for important confounders [12]; If the balance can not be achieved on very important confounders, then it is better to revise the design to account for such imbalance. Rubin (2001) has suggested three basic distributional conditions which must be simultaneously satisfied in a well balanced data. (1) The mean propensity score in the two groups being compared should be similar, e.g. the difference between means should be less than half of a standard deviation; (2) The variance of the propensity score in the two groups should be similar, e.g. less than ½ or greater than 2 are too extreme; (3) The variances of the residuals of the covariates after adjusting for the propensity score should be similar, e.g. less than ½ or greater than 2 are too extreme [12].

Here we are interested in the second utility of the propensity scores, namely in the properties of the distribution of propensity scores to predict “trustworthiness” or accuracy of the effect of exposure estimated from a given data. Two out of the three criteria presented by Rubin (2001) reflect direct discrepancies between the distributions of propensity scores for the two exposure levels. Indeed the degree of inequality of the distribution of the propensity scores for the exposure level is naturally related to the degree of association between the levels of exposure and other covariates, and hence affects the accuracy of the estimation exposure effect. One of the limitation of the Rubin's criteria is their suitable mostly for normal distributions.

## 1.3 AREA UNDER ROC CURVE

Receiver Operating Characteristic (ROC) curve is a plot of "sensitivity" versus "false positive rate" (or fraction) [14]. In general, the area under the ROC curve (AUC) computed for a predictive model is often termed as a measure of overall predictive accuracy, or discriminative ability of the model. It describes how well the predicted probabilities from the binary model, typically logistic regression model, classify patients into their actual class (e.g. exposed or non-exposed) or discriminate patients from the two different classes [14]. Here we will consider the AUC of the propensity model.

Theoretical value of the AUC could range from 0 to 1, corresponding to the cases when exposed subjects have propensity to be exposure always less (0) or always greater (1) than non-exposed subjects. However, it can be immediately seen that for the propensity scores or for any reasonable predictive model or diagnostic test the AUC does not assume values below 0.5 since if that had been the case, it would have violated the definition of the predictive probability or propensity score(a simple switching labels "exposed", "non-exposed" would produce a reasonable system with AUC greater than 0.5). Thus, in reasonable scenarios when the propensity score  is independent from the level of exposure, the true AUC is 0.5 and when the distributions of the propensity scores for the two levels of exposure has no overlap, the true AUC is 1.  For the logistic models Hosmer and Lemeshow suggest a general rule: (1) 0.5≤AUC<0.7 suggests poor discrimination; (2) 0.7≤AUC<0.8 suggests an acceptable discrimination; (3) 0.8≤AUC<0.9 suggests an excellent discrimination; (4) AUC≥0.9 suggests outstanding discrimination [6].

Being an estimate of a stochastic dominance of one of the distributions of propensity scores, AUC from the propensity model is a reflection of the covariates distribution among the exposure levels, hence, it reflects the degree of the association between covariates and exposure of interest. Generally speaking, the closer AUC is to 0.5the less strong the association is. The AUC of greater than 0.9 usually suggests complete nonoverlapping of the covariates across exposure groups and hence a strong association between the exposure of interest and other covariates.

The criteria suggested by Rubin, which reflect two specific differences between the distributions of the propensity scores, are most appropriate for the normally distributed data. For not necessarily normally distributed data another measure of the differences between the distributions is often used. This measure is the Wilcoxon statistic or a summary of the stochastic dominance of one of the distributions of the propensity scores, say corresponding to the exposed subjects, over the other distribution. In terms of the Receiver Operating Characteristic analysis this measure is equivalent to the area under the ROC curve. Although the AUC or c-statistic has been considered as an important index to report with the propensity model (typically logistic regression) [6, 15] to our knowledge it has not been used to characterize the trustworthiness of the estimated effect of exposure on the outcome of interest.

In relationship to the true propensity scores which are continuous (probability of a ties is zero), the area under the ROC curve can be interpreted as a probability that for a randomly selected exposed subject the propensity score would be higher than for a randomly selected unexposed subject. In the case when the ties in the propensity score are possible, the AUC is the above probability plus half of the probability of a tie – reflecting the principle that a forced binary (two exposure levels) discrimination between the two subjects with the same propensity score has a 1/2 chance to be correct.

Thus, area under ROC curve (AUC) provides a scalar statistics which quantifies the difference in the distributions of the propensity scores and has a potential to be an important predictor of the ability to obtain trustworthy inferences about effect of exposure with a given data. In this study we carry out a preliminary investigation of the relationship between the AUC of the propensity scores for exposure and the accuracy of the estimated effect of the exposure on the outcome of interest.

# 2.0 METHODS

The "main" model considered in this work is a logistic regression with outcome, y, as a dependent variable, and "independent" variables including the exposure of interest, x, and other covariates arranged in the vector **z**. We assume that there is no interaction between the exposure and other covariates.

$$\mathrm{Y|X,\ Z \sim bin(1,\ p_{y|xz})}$$

$$\log\left(\frac{p_{Y|X,Z}}{1-p_{Y|X,Z}}\right)=\alpha+\beta_{Y|X}*x+\beta_{Y|Z}^{1}*z_1+\ldots+\beta_{Y|Z}^{k}*z_k=a+\beta_{Y|X}\times x+\boldsymbol{\beta}_{\boldsymbol{Y|Z}}`\times\boldsymbol{z}$$

$$\boldsymbol{\beta}_{\boldsymbol{Y|Z}}=\begin{pmatrix}\beta_{Y|Z}^{1}\\ \vdots\\ \beta_{Y|Z}^{k}\end{pmatrix}\quad\boldsymbol{z}=\begin{pmatrix}z_1\\ \vdots\\ z_k\end{pmatrix}\qquad(1)$$

As a propensity model we consider a logistic regression with the exposure of interest, x, as dependent variable, the other covariates, **z**, as independent variables.

$$\mathrm{X|Z\sim bin\ (1,\ p_{x|z})}$$

$$\log\left(\frac{p_{X|Z}}{1-p_{X|Z}}\right)=\alpha+\beta_{X|Z}^{1}*z_1+\beta_{X|Z}^{k}*z_2+\ldots+\beta_k*z_k=\alpha+\boldsymbol{\beta}_{\boldsymbol{X|Z}}`\times\boldsymbol{z}$$

$$\boldsymbol{\beta}_{\boldsymbol{X|Z}}=\begin{pmatrix}\beta_{X|Z}^{1}\\ \vdots\\ \beta_{X|Z}^{k}\end{pmatrix}\quad\boldsymbol{z}=\begin{pmatrix}z_1\\ \vdots\\ z_k\end{pmatrix}\qquad(2)$$

The propensity score (Rosenbaum and Rubin, 1983) is the probability to be exposed given the values of the other observed covariates, i.e.

$$e(Z) = P(X = 1 \mid Z) = p_{x|z} = \left(1 + e^{-\left(\alpha + \boldsymbol{\beta}_{X|Z}` \times z\right)}\right)^{-1} \tag{3}$$

Being dependent on the random Z, the propensity score has its own probability distribution. The distribution of the propensity scores among the exposed (x=1) and unexposed (x=0) people are denoted as follows:

$$e(Z)|_{X=0} = e^{0}(Z) \sim F^{0} \quad e(Z)|_{X=1} = e^{1}(Z) \sim F^{1} \tag{4}$$

Then, the ability of the propensity score to discriminate between the two exposure levels can be comprehensively characterized by the ROC curve, which is the plot of True Positive Fraction (TPF, or sensitivity) versus False Positive Fraction (FPF, or 1-specificity), where the TPF and FPF are defined as follows:

$$\begin{aligned} FPF_{z|x}(t) &= P\left[e(Z) > t \mid X = 0\right] = P\left[e^{0}(Z) > t\right] = 1 - F^{0}(t) \\ TPF_{z|x}(t) &= P\left[e(Z) > t \mid X = 1\right] = P\left[e^{1}(Z) > t\right] = 1 - F^{1}(t) \end{aligned} \tag{5}$$

where t is a threshold which can be used to partition subjects into the two groups according to their propensity scores. When the propensity score is continuous (no ties possible) the area under the ROC curve can be written as:

$$A = \int TPF(t) dFPF(t) = P\left[e^{0} < e^{1}\right] \tag{6}$$

For non-continuous distribution of the propensity scores the expression becomes more complicated, i.e.:

$$A = P\left[e^{0} < e^{1}\right] + 0.5 \times P\left[e^{0} = e^{1}\right] \tag{7}$$

For the case where the propensity score assumes only two values (e.g. z is a single binary covariate), there is only one diagnostic threshold t which allows for nontrivial dichotomization of the propensity scores (trivial dichotomizations assign empty set to one of the groups). In this case we can drop the argument, t, for the True Positive and False Positive Fractions. The ROC curve, in this case, consists of two straight-line segments connecting points (0, 0), ($FPF_{z|x}$, $TPF_{z|x}$) and (1, 1) respectively; and we can express AUC in a more simple form, i.e.:

$$A_{z|x} = \frac{TPF_{z|x} + 1 - FPF_{z|x}}{2}, \tag{8}$$

where

$$TPF_{z|x}=P(z=1|x=1) \quad FPF_{z|x}=P(z=1|x=0) \tag{9}$$

The estimate of the AUC can be obtained using empirical estimator of the $FPF_{z|x}$ and $TPF_{z|x}$, or equivalently as a proportion of the times the estimated propensity score for exposed people was higher than that for unexposed, plus half of the proportion when the propensity scores are equal) i.e:

$$A = \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\psi\left(\hat{e}^0,\hat{e}^1\right)}{n_0 \times n_1} \quad where \quad \psi\left(\hat{e}^0,\hat{e}^1\right) = \begin{cases} 1 & \hat{e}^0 < \hat{e}^1 \\ 0.5 & \hat{e}^0 = \hat{e}^1 \\ 0 & \hat{e}^0 > \hat{e}^1 \end{cases} \tag{10}$$

where $n_0$, $n_1$ are the number of unexposed and exposed subjects correspondingly.

As a primary measure of the accuracy of the estimated effect of exposure in this work we use a mean squared error which is defined as follows:

$$MSE\left(\hat{\beta}_{Y|X}\right) = E\left[\left(\hat{\beta}_{Y|X} - \beta_{Y|X}\right)^2\right] \tag{11}$$

From the data the MSE was estimated according to the following expression:

$$\hat{MSE} = \left(\sum_{i=1}^{n}\frac{\beta_{Y|X} - \hat{\beta}_{Y|X}}{m}\right)^2 + \frac{m-1}{m}\hat{Var}\left(\hat{\beta}_{Y|X}\right) \tag{12}$$

m: number of simulations with converged logistic regression

$\beta_{Y|X}$: the true value of coefficient in the logistic regression (fixed and used to determine the distribution for generating the observations)

$\hat{\beta}_{Y|X}$: the estimated value of the coefficient in the logistic regression (using a dataset generated from the specified distribution)

The investigation was performed using a Monte Carlo simulation study. The simulation algorithm is shown in Figure 1. The possible scenarios when fitting the Propensity Model and Main Model are shown in Figure 2.

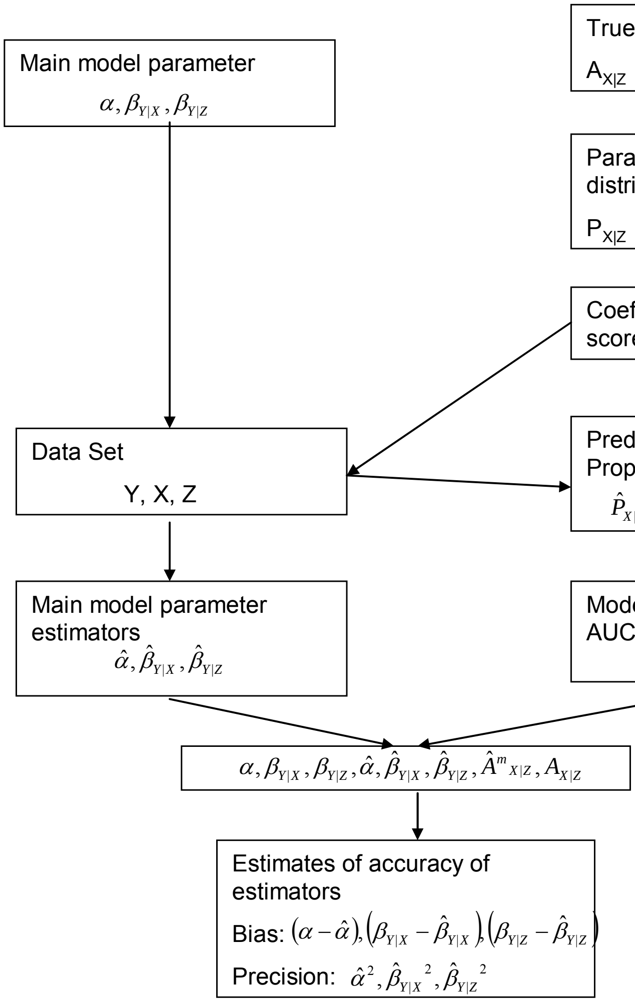
True AUC (X-Z)
$A_{X|Z}$
Main model parameter
$\alpha, \beta_{Y|X}, \beta_{Y|Z}$
Parameters of joint distribution of x and z
$P_{X|Z}$
Coefficients of propensity score model
Data Set
Y, X, Z
Predicted probability from Propensity model
$\hat{P}_{X|Z}$
Main model parameter estimators
$\hat{\alpha}, \hat{\beta}_{Y|X}, \hat{\beta}_{Y|Z}$
Model-based Estimated $\text{AUC}_{x|z}$ $\hat{A}^{m}{}_{X|Z}$
$\alpha, \beta_{Y|X}, \beta_{Y|Z}, \hat{\alpha}, \hat{\beta}_{Y|X}, \hat{\beta}_{Y|Z}, \hat{A}^{m}{}_{X|Z}, A_{X|Z}$
Estimates of accuracy of estimators
Bias: $(\alpha - \hat{\alpha}), (\beta_{Y|X} - \hat{\beta}_{Y|X}), (\beta_{Y|Z} - \hat{\beta}_{Y|Z})$
Precision: $\hat{\alpha}^2, \hat{\beta}_{Y|X}{}^2, \hat{\beta}_{Y|Z}{}^2$

**Figure 1 Simulation Algorithm**

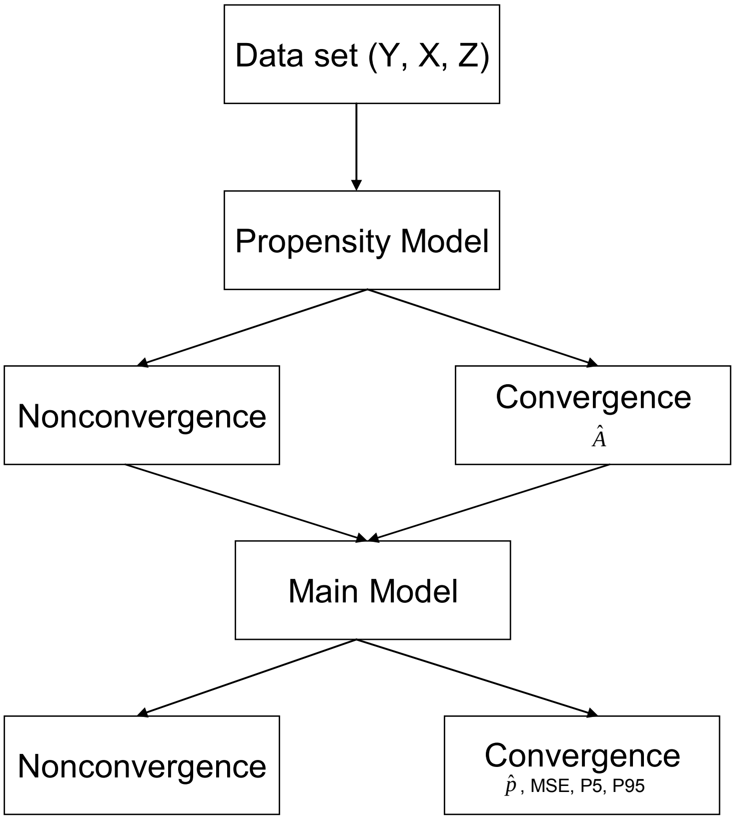

**Figure 2 Chart of Possible Scenario when Fitting the Two Models**

The considered values of the parameters of the propensity model and the prevalence of the exposure are listed in Table1. The considered values of the parameters of the main model are shown in Table 2.

The exposure of interest, x, was generated from a Bernoulli distributions with marginal probability of exposure (prevalence of exposure) $p_x$=P(x=1) of 0.2, 0.5 and 0.8. The degree of the association between the exposure, x, and the covariate, z, was determined by area under ROC curve (AUC) with values of 0.6, 0.7, 0.8, and 0.9 (which together with the prevalence of exposure determine the complete joint distribution of x and z). The values for $TPF_{z|x}$ and $FPF_{z|x}$ (eq. 9) shown in Table 1 were restricted to the following pattern: $TPF_{z|x}+FPF_{z|x}=1.1$. In our study the effect of exposure of interest on outcome ranged from moderate negative to moderate positive. Independently from the direction of the exposure effect, the effect of covariate on outcome also ranged from moderately negative to moderately positive with the absolute value similar that of the exposure effect. Thus, we model the scenario when the direction of the covariate effect was either same (enhancing the effect of the exposure in a univariate model) as or opposite (compensating for the effect of the exposure in a univariate model) to the effect of exposure of interest.

The sample size for a single simulated dataset was 50, 100, and 200 respectively. Peduzzi et al. found that standard asymptotic approximations are poor when sample size is smaller than ten times the number of parameter [17]. There are two parameters in our study, exposure of interest and one covariate. Therefore, a sample size of 50 was considered reasonable for this preliminary investigation (50>10*2). Finally, the total number of simulated dataset was 500.

**Table 1 Parameter Values of Propensity Model Considered in the Simulations**

| p(x=1) | AUC | p(z=1\|x=1) | $\beta_{x\|z}$ | α |
|---|---|---|---|---|
| 0.2 | 0.6 | 0.65 | 0.82 | -2.21 |
| 0.2 | 0.7 | 0.75 | 0.90 | -2.29 |
| 0.2 | 0.8 | 0.85 | 1.12 | -2.51 |
| 0.2 | 0.9 | 0.95 | 1.85 | -3.24 |
| 0.5 | 0.6 | 0.65 | 0.82 | -0.82 |
| 0.5 | 0.7 | 0.75 | 0.90 | -0.90 |
| 0.5 | 0.8 | 0.85 | 1.12 | -1.12 |
| 0.5 | 0.9 | 0.95 | 1.85 | -1.85 |
| 0.8 | 0.6 | 0.65 | 0.82 | 0.57 |
| 0.8 | 0.7 | 0.75 | 0.90 | 0.49 |
| 0.8 | 0.8 | 0.85 | 1.12 | 0.27 |
| 0.8 | 0.9 | 0.95 | 1.85 | -0.46 |

**Table 2 Parameter Values of Main Model Considered in the Simulations**

| $\beta_{Y\|X}$ | $\beta_{Y\|Z}$ | α |
|---|---|---|
| -1 | -0.95 | 0 |
| -1 | 0 | 0 |
| -1 | 0.95 | 0 |
| 0 | -0.05 | 0 |
| 0 | 0 | 0 |
| 0 | 0.05 | 0 |
| 1 | -1.05 | 0 |
| 1 | 0 | 0 |
| 1 | 1.05 | 0 |

The simulations were implemented using SAS v.9.1. The primary measure of interest in this study was mean square error (MSE), and the fifth and ninety-fifth percentile of the distribution of the estimated effect of exposure.

# 3.0 RESULTS

Table 3, 4 and 5 display the square root of the MSE (RMSE), and the 5th and 95th percentile of the empirical distribution of the estimator of the exposure effect for the sample size of 50, 100 and 200 correspondingly. The considered scenarios are indexed by specific values of the prevalence of the exposure ($p_x$), the AUC of the propensity model, and the parameters of the main model ($\beta_{Y|X}$, $\beta_{Y|Z}$).

The square root of MSE (RMSE) can be interpreted as the absolute distance between the true parameter value and the estimated parameter value. As AUC increases from 0.6 to 0.9, the RMSE increases for all considered values of other parameters (Tables 3, 4, 5). This trend can be seen clearly when plotting the square root of the MSE versus AUC (Figure 3, 4, 5). In the tables the relevance of the observed magnitude of the RMSE is illustrated with the ratio between the most extreme among not unlikely values (between 5th and 95th percentile) of the estimated odds ratio and the true odds ratio of the exposure. For example, for the sample size of 50, and AUC of 0.6 the estimate of the exposure effect is not unlikely to be as high as 13 times greater than the true effect, and when AUC increases to 0.8 this ratio become as high as 18. Naturally, the increasing sample size decreases the magnitude of the error and alleviates the effect of increasing AUC, for instance, the increase of the sample size to 200 does limit the not unlikely over- or under-estimated effect of exposure to being from 6 to 9 times (compared to 13-18) far from the true value depending on the AUC.

The results in Table 3, 4, 5 also demonstrate that the square root of MSE is the lowest when the prevalence of exposure is 0.5 and it increases when the prevalence changes away from 0.5 (to 0.2 or to 0.8). For example, for the prevalence of the exposure of 0.5 not unlikely estimates of the exposure effect can be up to 10.58 times lower/greater than the true effect, while for the prevalence of 0.2 or 0.8 it can be up to 14.29 times lower/greater (Table 4). The prevalence does not seem to substantially modify the effect of AUC on RMSE (Figure 6).

The size of a sample has a significant effect on the RMSE on average but does not seem to substantially modify the effect of AUC on RMSE (Figure 8). On average the RMSE decreases as the sample size changes from 50 to 100, 200. Not unlikely estimates of the exposure effect can be up to 20 times far from the truth for the sample size 50 and up to 8.3 times for sample size of 200. However, for all considered  sample sizes, the AUC increasing from 0.6 to 0.9 increases the maximum ratio between the true and not unlikely estimates of the odds ratio by approximately 1.5 (20/13 for 50, 11/7 for 100, 9/6 for 200).

In summary, the results of the conducted simulation study indicate that the inaccuracy of the estimated effect of the exposure substantially increases with the increasing area under the ROC curve (AUC). The rate of change of the RMSE is the highest for AUC>0.8 and the lowest for AUCs<0.7. Other parameters, such as $p_x$, $\beta_{Y|X}$, $\beta_{Y|Z}$ and sample size, do not exert substantial effect on the general shape of the observed trend. However, the height of the trend curves (or average RMSE) is affected by both sample size and prevalence of exposure.

**Table 3 Square root of MSE, $5^{th}$ and $95^{th}$ percentile of the estimated exposure effect (sample size=50)**

| | | | AUC=0.6 | | | | AUC=0.7 | | | | AUC=0.8 | | | | AUC=0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| px | by_x | by_z | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio |
| 0.2 | -1 | -0.95 | 0.857 | -1.106 | **1.588** | 13.30 | 0.898 | -1.043 | **1.699** | 14.86 | 0.98 | -1.187 | **1.906** | 18.28 | 0.974 | -1.332 | **1.74** | 15.48 |
| 0.2 | -1 | 0 | 0.843 | -1.389 | **1.47** | 11.82 | 0.853 | -1.377 | **1.386** | 10.87 | 0.955 | -1.579 | **1.574** | 13.11 | 1.015 | -1.708 | **1.801** | 16.46 |
| 0.2 | -1 | 0.95 | 0.883 | -1.612 | **1.446** | 11.54 | 0.87 | -1.518 | **1.357** | 10.55 | 0.94 | -1.596 | **1.623** | 13.77 | 0.953 | -1.325 | **1.693** | 14.77 |
| 0.2 | 0 | -0.05 | 0.842 | -1.354 | **1.44** | 4.22 | 0.822 | -1.269 | **1.504** | 4.50 | 0.97 | -1.584 | **1.694** | 5.44 | 1.067 | **-1.832** | 1.56 | 6.25^(-1) |
| 0.2 | 0 | 0 | 0.84 | -1.356 | **1.44** | 4.22 | 0.822 | -1.248 | **1.516** | 4.55 | 0.974 | -1.63 | **1.675** | 5.34 | 1.054 | **-1.792** | 1.56 | 5.88^(-1) |
| 0.2 | 0 | 0.05 | 0.835 | -1.354 | **1.393** | 4.03 | 0.829 | -1.237 | **1.516** | 4.56 | 0.968 | -1.606 | **1.714** | 5.55 | 1.042 | **-1.792** | 1.56 | 5.88^(-1) |
| 0.2 | 1 | -1.05 | 0.905 | **-1.341** | 1.781 | 10^(-1) | 0.899 | **-1.32** | 1.708 | 10^(-1) | 0.966 | **-1.56** | 1.718 | 12.5^(-1) | 1.05 | **-1.931** | 1.485 | 20^(-1) |
| 0.2 | 1 | 0 | 0.811 | **-1.274** | 1.328 | 10^(-1) | 0.843 | **-1.186** | 1.484 | 9.09^(-1) | 1.003 | **-1.537** | 1.746 | 12.5^(-1) | 1.031 | **-1.827** | 1.708 | 16.67^(-1) |
| 0.2 | 1 | 1.05 | 0.847 | **-1.494** | 1.066 | 12.5^(-1) | 0.848 | **-1.603** | 0.945 | 14.29^(-1) | 0.901 | **-1.693** | 1.041 | 14.29^(-1) | 0.881 | **-1.602** | 1.079 | 14.29^(-1) |
| 0.5 | -1 | -0.95 | 0.788 | -1.459 | **1.141** | 8.51 | 0.87 | -1.436 | **1.352** | 10.51 | 0.935 | -1.51 | **1.542** | 12.70 | 1.051 | -1.944 | **1.436** | 11.42 |
| 0.5 | -1 | 0 | 0.718 | -1.311 | **1** | 7.39 | 0.779 | -1.392 | **1.174** | 8.79 | 0.854 | -1.416 | **1.442** | 11.49 | 0.921 | -1.456 | **1.598** | 13.43 |
| 0.5 | -1 | 0.95 | 0.668 | -1.342 | **0.875** | 6.52 | 0.73 | -1.434 | **1.023** | 7.56 | 0.76 | -1.344 | **1.188** | 8.91 | 0.809 | -0.902 | **1.575** | 13.13 |
| 0.5 | 0 | -0.05 | 0.608 | **-1.049** | 0.966 | 2.86^(-1) | 0.718 | **-1.18** | 1.176 | 3.23^(-1) | 0.843 | -1.438 | **1.491** | 4.44 | 0.859 | -1.405 | **1.478** | 4.38 |
| 0.5 | 0 | 0 | 0.61 | **-1.059** | 0.977 | 2.86^(-1) | 0.721 | -1.187 | **1.195** | 3.30 | 0.843 | -1.429 | **1.471** | 4.35 | 0.878 | -1.39 | **1.5** | 4.48 |
| 0.5 | 0 | 0.05 | 0.604 | **-1.044** | 0.972 | 2.86^(-1) | 0.717 | -1.175 | **1.229** | 3.42 | 0.839 | **-1.387** | 1.356 | 4^(-1) | 0.872 | -1.367 | **1.478** | 4.38 |
| 0.5 | 1 | -1.05 | 0.667 | **-1.027** | 1.251 | 7.69^(-1) | 0.757 | **-1.043** | 1.506 | 7.69^(-1) | 0.797 | **-1.244** | 1.307 | 9.09^(-1) | 0.868 | **-1.744** | 0.941 | 16.67^(-1) |
| 0.5 | 1 | 0 | 0.665 | **-0.99** | 1.185 | 7.14^(-1) | 0.762 | **-1.139** | 1.412 | 8.33^(-1) | 0.876 | **-1.432** | 1.487 | 11.11^(-1) | 0.909 | **-1.436** | 1.519 | 11.11^(-1) |
| 0.5 | 1 | 1.05 | 0.722 | **-1.026** | 1.354 | 7.69^(-1) | 0.8 | **-1.254** | 1.495 | 10^(-1) | 0.947 | **-1.531** | 1.625 | 12.5^(-1) | 1.002 | **-1.447** | 1.755 | 11.11^(-1) |
| 0.8 | -1 | -0.95 | 0.968 | -1.788 | **1.379** | 10.79 | 0.956 | -1.744 | **1.525** | 12.49 | 1.03 | -1.726 | **1.768** | 15.92 | 1.072 | -1.702 | **1.818** | 16.74 |
| 0.8 | -1 | 0 | 0.932 | -1.773 | **1.247** | 9.46 | 0.953 | -1.707 | **1.441** | 11.48 | 0.958 | -1.621 | **1.654** | 14.20 | 0.974 | -1.154 | **1.915** | 18.44 |
| 0.8 | -1 | 0.95 | 0.897 | -1.867 | **1.161** | 8.68 | 0.9 | -1.674 | **1.28** | 9.77 | 0.879 | -1.397 | **1.41** | 11.13 | 0.965 | -0.966 | **1.981** | 19.70 |
| 0.8 | 0 | -0.05 | 0.87 | **-1.546** | 1.325 | 4.76^(-1) | 0.908 | **-1.557** | 1.487 | 4.76^(-1) | 0.963 | **-1.679** | 1.571 | 5.26^(-1) | 0.937 | -1.438 | **1.65** | 5.21 |
| 0.8 | 0 | 0 | 0.874 | **-1.537** | 1.325 | 4.55^(-1) | 0.909 | **-1.629** | 1.403 | 5^(-1) | 0.955 | **-1.679** | 1.571 | 5.26^(-1) | 0.933 | -1.438 | **1.674** | 5.33 |
| 0.8 | 0 | 0.05 | 0.877 | **-1.537** | 1.287 | 4.55^(-1) | 0.909 | **-1.581** | 1.435 | 4.76^(-1) | 0.959 | **-1.693** | 1.594 | 5.56^(-1) | 0.929 | -1.453 | **1.636** | 5.14 |
| 0.8 | 1 | -1.05 | 0.869 | **-1.375** | 1.529 | 11.11^(-1) | 0.863 | **-1.323** | 1.41 | 10^(-1) | 0.923 | **-1.677** | 1.37 | 14.29^(-1) | 0.978 | **-1.833** | 0.827 | 16.67^(-1) |
| 0.8 | 1 | 0 | 0.872 | **-1.362** | 1.474 | 11.11^(-1) | 0.904 | **-1.444** | 1.6 | 11.11^(-1) | 0.952 | **-1.679** | 1.443 | 14.29^(-1) | 0.995 | **-2.064** | 1.168 | 20^(-1) |
| 0.8 | 1 | 1.05 | 0.932 | **-1.459** | 1.637 | 11.11^(-1) | 0.965 | **-1.615** | 1.672 | 14.29^(-1) | 1.072 | **-1.752** | 1.724 | 16.67^(-1) | 1.125 | **-2.01** | 1.883 | 20^(-1) |

*: True OR/OR $5^{th}$, or True OR/OR $95^{th}$, whichever is farther from 1.

**Table 4 Square root of MSE, 5th and 95th percentile of the estimated exposure effect (sample size=100)**

| px | by_x | by_z | AUC=0.6 | | | | AUC=0.7 | | | | AUC=0.8 | | | | AUC=0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio |
| 0.2 | -1 | -0.95 | 0.667 | -1.223 | **0.971** | 7.18 | 0.676 | -1.256 | **0.958** | 7.08 | 0.711 | -1.262 | **1.079** | 7.99 | 0.853 | -1.398 | **1.403** | 11.06 |
| 0.2 | -1 | 0 | 0.625 | -1.281 | **0.906** | 6.73 | 0.616 | -1.113 | **0.913** | 6.77 | 0.712 | -1.24 | **1.038** | 7.67 | 0.811 | -1.565 | **1.182** | 8.86 |
| 0.2 | -1 | 0.95 | 0.571 | -1.033 | **0.863** | 6.44 | 0.574 | -0.956 | **0.913** | 6.77 | 0.684 | -1.289 | **0.952** | 7.04 | 0.798 | -1.485 | **1.117** | 8.30 |
| 0.2 | 0 | -0.05 | 0.549 | **-0.912** | 0.851 | 2.5^(-1) | 0.556 | -0.867 | **0.919** | 2.51 | 0.643 | **-1.044** | 1.028 | 2.86^(-1) | 0.797 | **-1.416** | 1.273 | 4.17^(-1) |
| 0.2 | 0 | 0 | 0.542 | **-0.914** | 0.834 | 2.5^(-1) | 0.552 | -0.9 | **0.944** | 2.57 | 0.636 | **-1.037** | 1.003 | 2.78^(-1) | 0.778 | **-1.416** | 1.18 | 4.17^(-1) |
| 0.2 | 0 | 0.05 | 0.535 | **-0.914** | 0.838 | 2.5^(-1) | 0.552 | -0.83 | **0.948** | 2.58 | 0.636 | -0.999 | **1.036** | 2.82 | 0.781 | **-1.361** | 1.198 | 3.85^(-1) |
| 0.2 | 1 | -1.05 | 0.585 | **-0.919** | 0.997 | 6.67^(-1) | 0.599 | **-0.895** | 1.084 | 6.67^(-1) | 0.719 | **-0.965** | 1.305 | 7.14^(-1) | 0.791 | **-1.201** | 1.398 | 9.09^(-1) |
| 0.2 | 1 | 0 | 0.619 | **-0.883** | 1.2 | 6.67^(-1) | 0.642 | **-0.877** | 1.279 | 6.67^(-1) | 0.71 | **-0.993** | 1.412 | 7.14^(-1) | 0.806 | **-1.227** | 1.398 | 9.09^(-1) |
| 0.2 | 1 | 1.05 | 0.703 | **-1.077** | 1.301 | 7.69^(-1) | 0.707 | **-1.065** | 1.359 | 7.69^(-1) | 0.77 | **-1.16** | 1.325 | 8.33^(-1) | 0.884 | **-1.603** | 1.351 | 14.29^(-1) |
| 0.5 | -1 | -0.95 | 0.528 | -0.961 | **0.803** | 6.07 | 0.556 | -0.978 | **0.891** | 6.62 | 0.63 | -1.074 | **1.046** | 7.74 | 0.85 | -1.505 | **1.36** | 10.58 |
| 0.5 | -1 | 0 | 0.493 | -0.875 | **0.708** | 5.52 | 0.526 | -0.977 | **0.702** | 5.48 | 0.598 | -1.062 | **0.904** | 6.71 | 0.787 | -1.327 | **1.212** | 9.13 |
| 0.5 | -1 | 0.95 | 0.481 | -0.879 | **0.669** | 5.31 | 0.53 | -0.982 | **0.645** | 5.18 | 0.596 | -1.156 | **0.82** | 6.17 | 0.688 | -1.234 | **1.037** | 7.67 |
| 0.5 | 0 | -0.05 | 0.437 | **-0.722** | 0.69 | 2.04^(-1) | 0.489 | **-0.817** | 0.711 | 2.27^(-1) | 0.577 | **-0.969** | 0.926 | 2.63^(-1) | 0.761 | **-1.345** | 1.253 | 3.85^(-1) |
| 0.5 | 0 | 0 | 0.436 | **-0.733** | 0.691 | 2.08^(-1) | 0.49 | **-0.834** | 0.734 | 2.33^(-1) | 0.574 | **-0.938** | 0.909 | 2.56^(-1) | 0.766 | **-1.315** | 1.282 | 3.7^(-1) |
| 0.5 | 0 | 0.05 | 0.435 | **-0.721** | 0.695 | 2.04^(-1) | 0.489 | **-0.852** | 0.727 | 2.33^(-1) | 0.567 | **-0.947** | 0.922 | 2.56^(-1) | 0.768 | **-1.345** | 1.254 | 3.85^(-1) |
| 0.5 | 1 | -1.05 | 0.458 | **-0.728** | 0.755 | 6.25^(-1) | 0.511 | **-0.755** | 0.877 | 5.88^(-1) | 0.594 | **-0.885** | 1.04 | 6.67^(-1) | 0.718 | **-1.154** | 1.228 | 8.33^(-1) |
| 0.5 | 1 | 0 | 0.467 | **-0.74** | 0.834 | 6.25^(-1) | 0.516 | **-0.837** | 0.853 | 6.25^(-1) | 0.594 | **-0.889** | 0.997 | 6.67^(-1) | 0.769 | **-1.241** | 1.419 | 9.09^(-1) |
| 0.5 | 1 | 1.05 | 0.526 | **-0.818** | 0.928 | 6.25^(-1) | 0.58 | **-0.936** | 0.954 | 7.14^(-1) | 0.66 | **-1.105** | 1.049 | 8.33^(-1) | 0.838 | **-1.339** | 1.483 | 10^(-1) |
| 0.8 | -1 | -0.95 | 0.626 | -1.073 | **1.002** | 7.40 | 0.618 | -1.036 | **0.986** | 7.28 | 0.74 | -1.365 | **1.095** | 8.13 | 0.886 | -1.573 | **1.47** | 11.82 |
| 0.8 | -1 | 0 | 0.62 | -1.057 | **0.876** | 6.53 | 0.624 | -1.184 | **0.902** | 6.70 | 0.749 | -1.385 | **1.05** | 7.77 | 0.869 | -1.38 | **1.505** | 12.24 |
| 0.8 | -1 | 0.95 | 0.619 | -1.184 | **0.784** | 5.95 | 0.629 | -1.197 | **0.861** | 6.43 | 0.753 | -1.516 | **0.97** | 7.17 | 0.768 | -1.199 | **1.412** | 11.15 |
| 0.8 | 0 | -0.05 | 0.594 | **-1.043** | 0.879 | 2.86^(-1) | 0.6 | **-1.041** | 1.005 | 2.86^(-1) | 0.732 | **-1.303** | 1.16 | 3.70^(-1) | 0.944 | -1.588 | **1.747** | 5.74 |
| 0.8 | 0 | 0 | 0.596 | **-1.068** | 0.847 | 2.94^(-1) | 0.595 | **-1.031** | 0.944 | 2.78^(-1) | 0.728 | **-1.312** | 1.159 | 3.70^(-1) | 0.936 | -1.593 | **1.768** | 5.86 |
| 0.8 | 0 | 0.05 | 0.596 | **-1.077** | 0.864 | 2.94^(-1) | 0.595 | **-1.062** | 0.953 | 2.86^(-1) | 0.725 | **-1.306** | 1.117 | 3.70^(-1) | 0.933 | -1.533 | **1.768** | 5.86 |
| 0.8 | 1 | -1.05 | 0.617 | **-0.923** | 1.155 | 6.67^(-1) | 0.62 | **-0.939** | 1.152 | 7.14^(-1) | 0.739 | -1.129 | **1.376** | 1.46 | 0.841 | **-1.444** | 1.217 | 11.11^(-1) |
| 0.8 | 1 | 0 | 0.597 | **-0.974** | 1.017 | 7.14^(-1) | 0.622 | **-0.999** | 0.964 | 7.14^(-1) | 0.752 | -1.119 | **1.282** | 1.33 | 0.918 | **-1.63** | 1.493 | 14.29^(-1) |
| 0.8 | 1 | 1.05 | 0.674 | **-1.115** | 0.968 | 8.33^(-1) | 0.681 | **-1.213** | 1.147 | 9.09^(-1) | 0.8 | -1.359 | **1.243** | 1.27 | 1.023 | **-1.649** | 1.796 | 14.29^(-1) |

*: True OR/OR 5$^{th}$, or True OR/OR 95$^{th}$, whichever is farther from 1.

**Table 5 Square root of MSE, 5th and 95th percentile of the estimated exposure effect (sample size=200)**

| px | by_x | by_z | AUC=0.6 | | | | AUC=0.7 | | | | AUC=0.8 | | | | AUC=0.9 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio | $\sqrt{MSE}$ | P5 | P95 | *Ratio |
| 0.2 | -1 | -0.95 | 0.501 | -0.905 | **0.753** | 5.77 | 0.52 | -0.896 | **0.784** | 5.95 | 0.571 | -1.039 | **0.781** | 5.94 | 0.67 | -1.165 | **1.034** | 7.64 |
| 0.2 | -1 | 0 | 0.431 | -0.74 | **0.639** | 5.15 | 0.439 | -0.736 | **0.675** | 5.34 | 0.482 | -0.839 | **0.707** | 5.51 | 0.563 | -1.024 | **0.828** | 6.22 |
| 0.2 | -1 | 0.95 | 0.412 | -0.737 | **0.679** | 5.36 | 0.4 | -0.631 | **0.692** | 5.43 | 0.447 | -0.788 | **0.668** | 5.30 | 0.559 | -1.044 | **0.808** | 6.10 |
| 0.2 | 0 | -0.05 | 0.39 | -0.655 | 0.655 | 1.93 | 0.379 | -0.577 | **0.675** | 1.97 | 0.423 | -0.653 | **0.735** | 2.09 | 0.533 | -0.842 | **0.888** | 2.43 |
| 0.2 | 0 | 0 | 0.388 | **-0.648** | 0.642 | 1.92^(-1) | 0.378 | -0.579 | **0.685** | 1.98 | 0.418 | -0.653 | **0.721** | 2.06 | 0.524 | -0.829 | **0.853** | 2.35 |
| 0.2 | 0 | 0.05 | 0.382 | -0.625 | **0.639** | 1.89 | 0.37 | -0.565 | **0.654** | 1.92 | 0.412 | -0.632 | **0.704** | 2.02 | 0.521 | **-0.869** | 0.857 | 2.38^(-1) |
| 0.2 | 1 | -1.05 | 0.417 | **-0.674** | 0.698 | 5.26^(-1) | 0.433 | **-0.637** | 0.795 | 5.26^(-1) | 0.49 | **-0.691** | 0.922 | 5.56^(-1) | 0.594 | **-0.787** | 1.125 | 5.88^(-1) |
| 0.2 | 1 | 0 | 0.424 | **-0.628** | 0.775 | 5^(-1) | 0.453 | **-0.614** | 0.822 | 5^(-1) | 0.485 | **-0.681** | 0.821 | 5.26^(-1) | 0.547 | **-0.817** | 0.938 | 6.25^(-1) |
| 0.2 | 1 | 1.05 | 0.543 | **-0.753** | 1.054 | 5.88^(-1) | 0.598 | **-0.732** | 1.23 | 5.56^(-1) | 0.627 | **-0.83** | 1.262 | 6.25^(-1) | 0.706 | **-1.033** | 1.229 | 7.69^(-1) |
| 0.5 | -1 | -0.95 | 0.356 | -0.612 | **0.553** | 4.73 | 0.379 | -0.657 | **0.603** | 4.96 | 0.439 | -0.725 | **0.715** | 5.56 | 0.596 | -0.962 | **1.027** | 7.59 |
| 0.5 | -1 | 0 | 0.324 | -0.594 | **0.479** | 4.39 | 0.359 | -0.643 | **0.569** | 4.80 | 0.406 | -0.693 | **0.631** | 5.11 | 0.561 | -0.924 | **0.904** | 6.71 |
| 0.5 | -1 | 0.95 | 0.335 | -0.619 | **0.489** | 4.43 | 0.365 | -0.72 | **0.514** | 4.54 | 0.428 | -0.812 | **0.612** | 5.01 | 0.557 | -0.972 | **0.722** | 5.60 |
| 0.5 | 0 | -0.05 | 0.308 | **-0.529** | 0.488 | 1.69^(-1) | 0.34 | **-0.578** | 0.551 | 1.79^(-1) | 0.399 | -0.665 | **0.676** | 1.97 | 0.539 | -0.795 | **0.899** | 2.46 |
| 0.5 | 0 | 0 | 0.306 | **-0.546** | 0.474 | 1.72^(-1) | 0.335 | **-0.59** | 0.517 | 1.82^(-1) | 0.393 | -0.627 | **0.644** | 1.90 | 0.537 | -0.859 | **0.93** | 2.54 |
| 0.5 | 0 | 0.05 | 0.303 | **-0.529** | 0.466 | 1.69^(-1) | 0.331 | **-0.607** | 0.521 | 1.82^(-1) | 0.387 | -0.619 | **0.62** | 1.86 | 0.524 | **-0.901** | 0.866 | 2.44^(-1) |
| 0.5 | 1 | -1.05 | 0.326 | **-0.531** | 0.546 | 4.55^(-1) | 0.353 | **-0.592** | 0.619 | 5^(-1) | 0.436 | **-0.644** | 0.722 | 5.26^(-1) | 0.574 | **-0.795** | 1.046 | 5.88^(-1) |
| 0.5 | 1 | 0 | 0.325 | **-0.508** | 0.565 | 4.55^(-1) | 0.353 | **-0.586** | 0.583 | 5^(-1) | 0.413 | **-0.614** | 0.712 | 5^(-1) | 0.56 | **-0.83** | 0.92 | 6.25^(-1) |
| 0.5 | 1 | 1.05 | 0.378 | **-0.629** | 0.659 | 5^(-1) | 0.404 | **-0.653** | 0.695 | 5.26^(-1) | 0.468 | **-0.713** | 0.764 | 5.56^(-1) | 0.622 | **-1.013** | 0.908 | 7.69^(-1) |
| 0.8 | -1 | -0.95 | 0.419 | -0.696 | **0.668** | 5.30 | 0.432 | -0.731 | **0.684** | 5.38 | 0.494 | -0.838 | **0.806** | 6.08 | 0.697 | -1.155 | **1.153** | 8.61 |
| 0.8 | -1 | 0 | 0.404 | -0.675 | **0.628** | 5.09 | 0.417 | -0.69 | **0.602** | 4.96 | 0.489 | -0.784 | **0.835** | 6.26 | 0.695 | -1.245 | **1.029** | 7.60 |
| 0.8 | -1 | 0.95 | 0.415 | -0.745 | **0.563** | 4.77 | 0.442 | -0.823 | **0.621** | 5.05 | 0.523 | -0.982 | **0.721** | 5.59 | 0.665 | -1.187 | **1.016** | 7.51 |
| 0.8 | 0 | -0.05 | 0.399 | **-0.677** | 0.597 | 1.96^(-1) | 0.405 | **-0.687** | 0.639 | 2^(-1) | 0.478 | **-0.816** | 0.791 | 2.27^(-1) | 0.678 | **-1.233** | 1.075 | 3.45^(-1) |
| 0.8 | 0 | 0 | 0.399 | **-0.698** | 0.6 | 2^(-1) | 0.403 | **-0.717** | 0.632 | 2.04^(-1) | 0.476 | -0.792 | 0.792 | 2.22^(-1) | 0.681 | **-1.234** | 1.075 | 3.45^(-1) |
| 0.8 | 0 | 0.05 | 0.391 | **-0.684** | 0.583 | 1.96^(-1) | 0.401 | **-0.697** | 0.628 | 2^(-1) | 0.476 | -0.787 | **0.8** | 2.23 | 0.676 | **-1.215** | 1.077 | 3.33^(-1) |
| 0.8 | 1 | -1.05 | 0.401 | **-0.655** | 0.657 | 5.26^(-1) | 0.417 | **-0.656** | 0.699 | 5.26^(-1) | 0.493 | **-0.738** | 0.866 | 5.56^(-1) | 0.7 | **-0.98** | 1.337 | 7.14^(-1) |
| 0.8 | 1 | 0 | 0.39 | **-0.628** | 0.64 | 5^(-1) | 0.402 | **-0.668** | 0.678 | 5.26^(-1) | 0.47 | **-0.703** | 0.806 | 5.56^(-1) | 0.715 | **-1.108** | 1.297 | 8.33^(-1) |
| 0.8 | 1 | 1.05 | 0.435 | **-0.748** | 0.669 | 5.88^(-1) | 0.45 | **-0.803** | 0.688 | 6.25^(-1) | 0.508 | **-0.882** | 0.883 | 6.67^(-1) | 0.736 | **-1.159** | 1.211 | 8.33^(-1) |

*: True OR/OR 5$^{th}$, or True OR/OR 95$^{th}$, whichever is farther from 1.

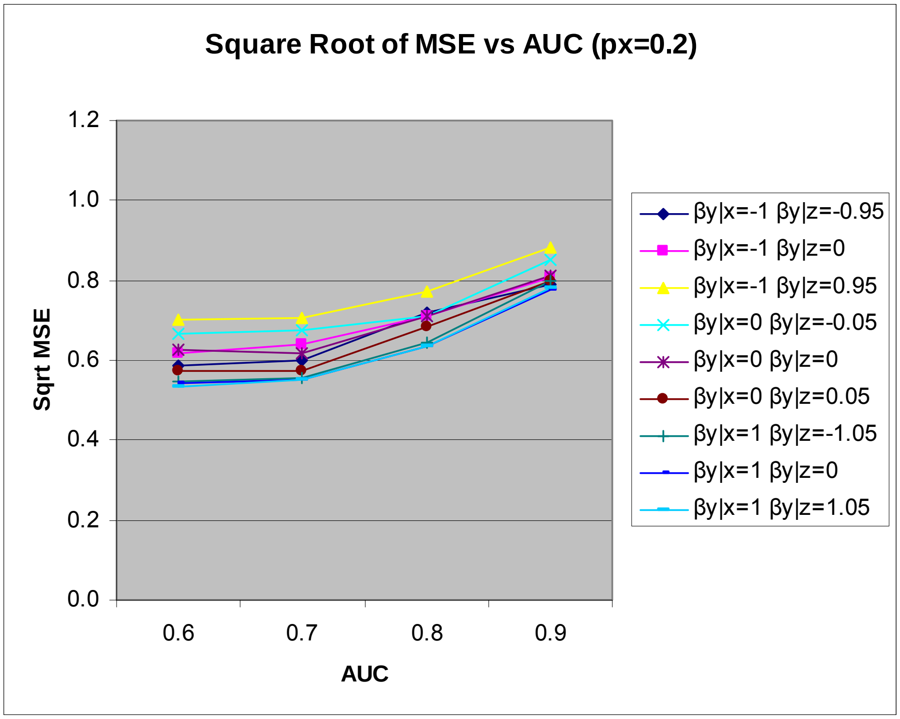

**Figure 3 Square Root of MSE vs. AUC (exposure prevalence $p_x$= 0.2, sample size=100)**

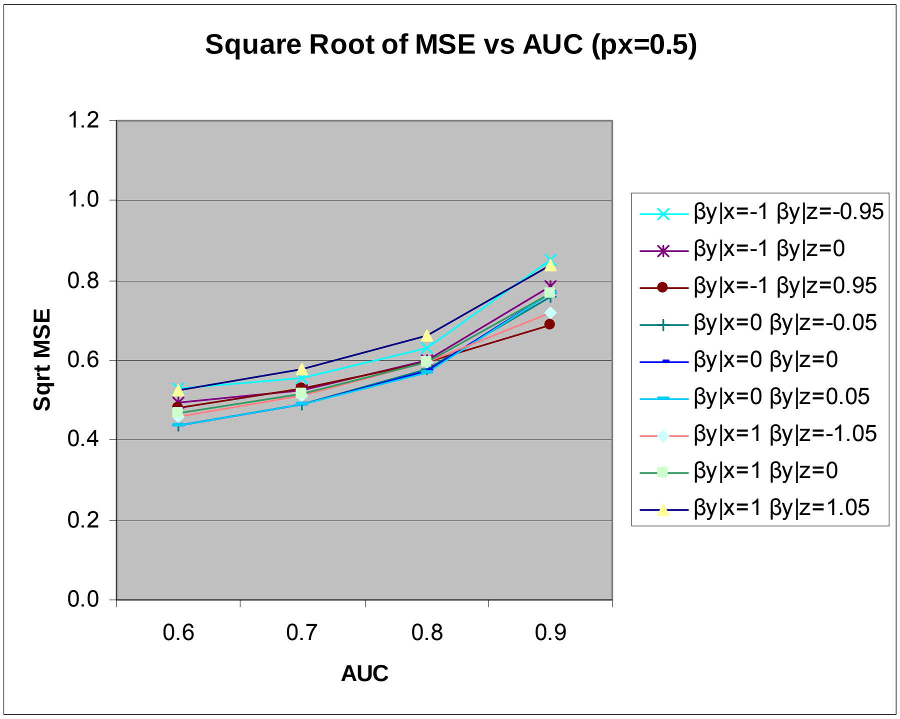

**Figure 4 Square Root of MSE vs. AUC (exposure prevalence $p_x$= 0.5, sample size=100)**

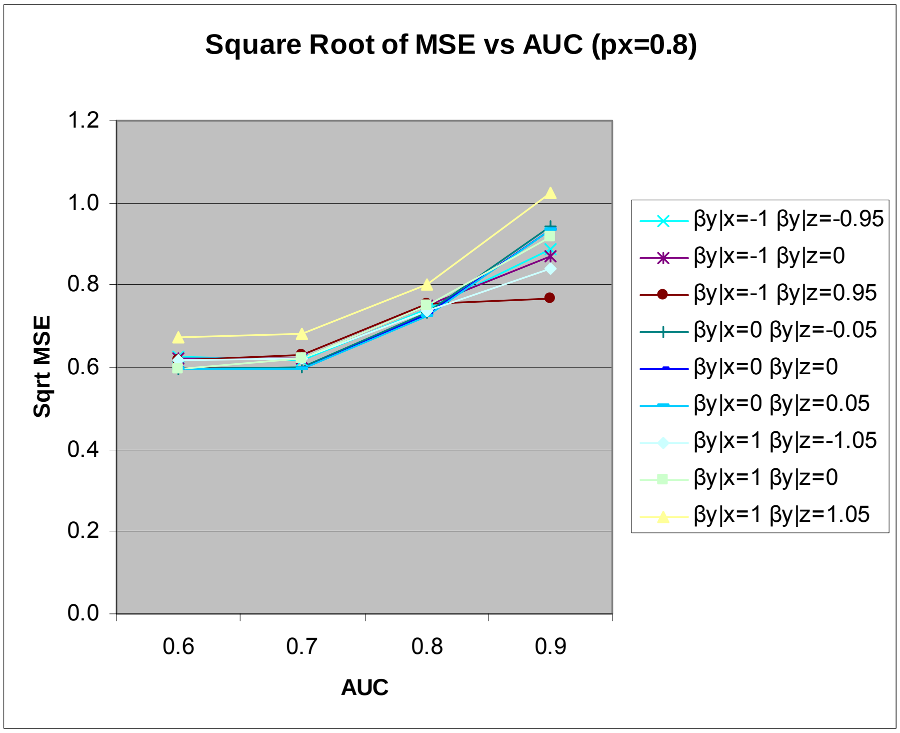

**Figure 5 Square Root of MSE vs. AUC (exposure prevalence $p_x$=0.8, sample size=100)**

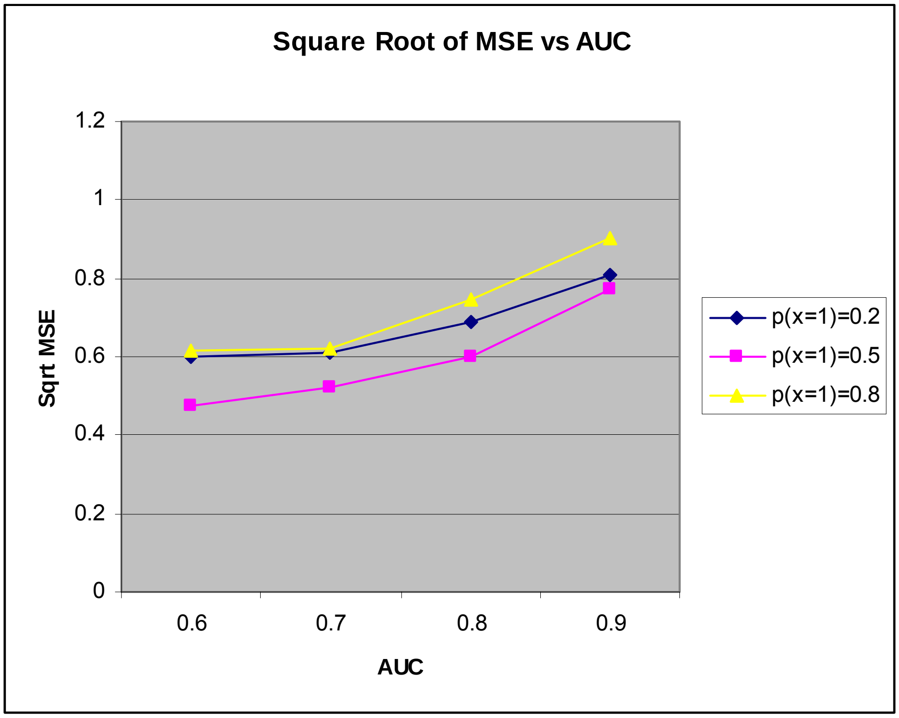

**Figure 6 Square Root of MSE vs. AUC (sample size = 100)**

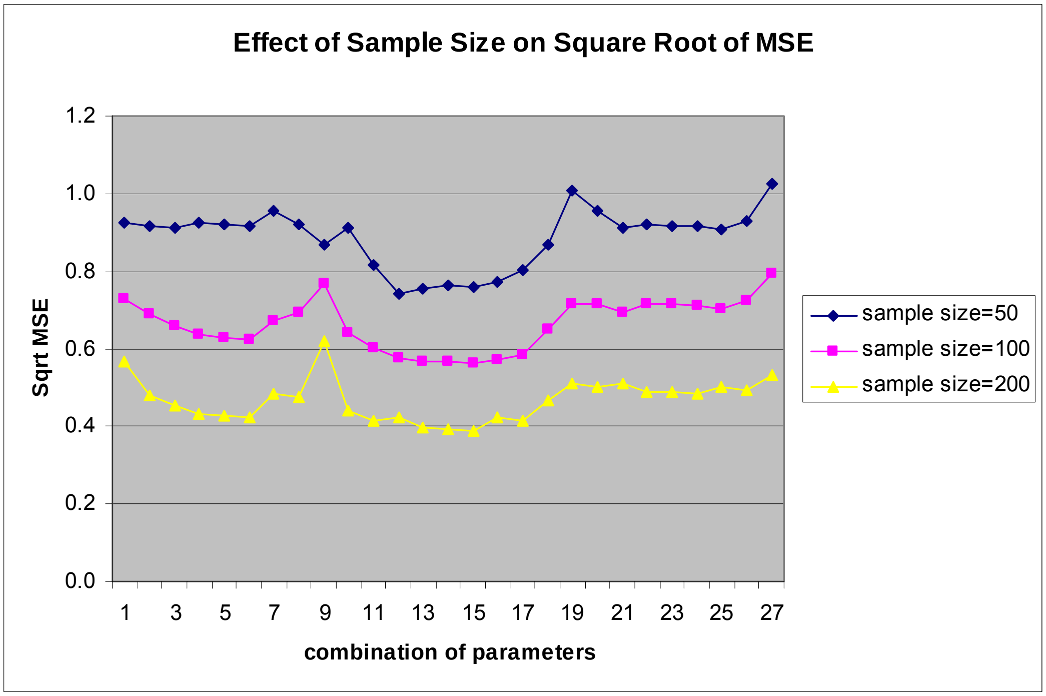

**Figure 7 Effect of Sample Size on Square Root of MSE**

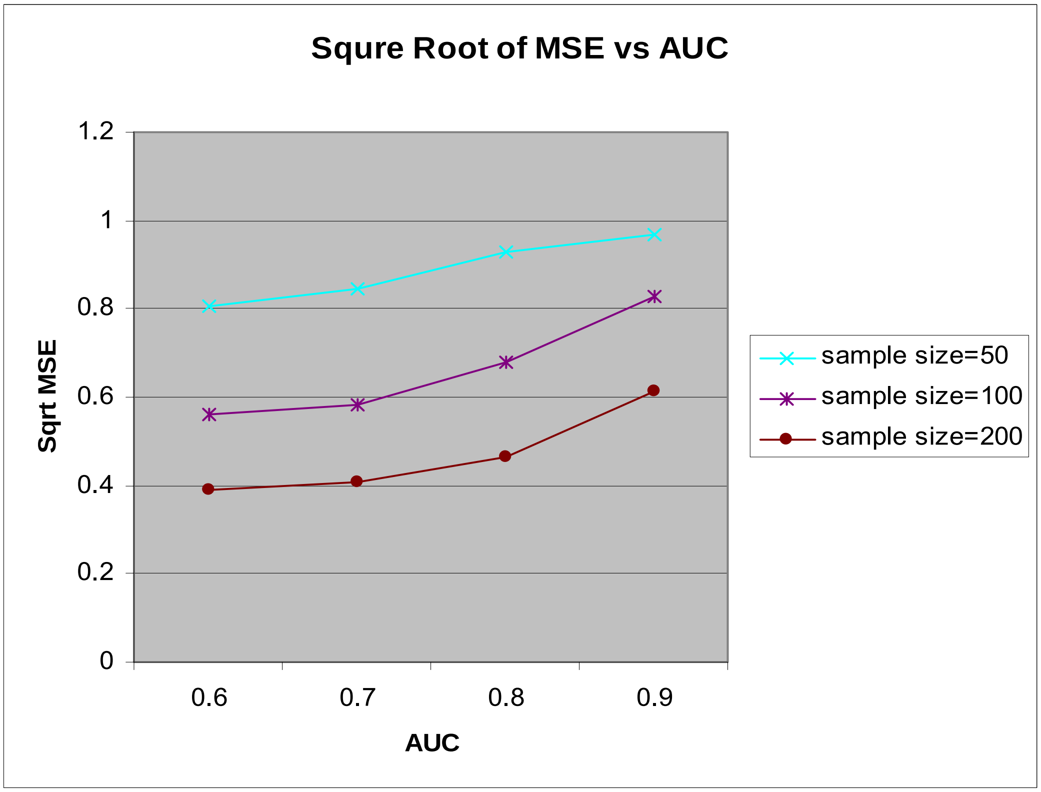

**Figure 8 Square Root of MSE vs. AUC**

# 4.0 DISCUSSION

We conducted a Monte Carlo simulation study to investigate the relationship between the area under the ROC curve (AUC) of the propensity model for the exposure and the estimated effect of exposure on the outcome of interest. We demonstrated that AUC provides a single, a relatively easy to compute, and suitable for various kinds of data statistic which could be used as an important indicator of the trustworthiness of the estimated effect of exposure. Our simulation results indicate that when AUC of the propensity model is larger than 0.8 the estimated effect of the exposure is highly inaccurate, at the same time the AUCs of less than 0.7 are not associated with any substantial increase of the inaccuracy.

Because of the limitations of the conducted simulation study the above recommendations have only a preliminary nature. In a real-world medical research, there are almost always multiple risk factors for the outcome of interest, thus multiple covariates need to be included with the exposure of interest in the model. Thus, the simple case of a single binary covariate considered in this work merely enables an identification of some of the trends which, if later confirmed to be general, may help in developing a more detailed procedure for using AUC of the propensity scores for diagnoses of the potential trustworthiness of the estimated exposure effect.

In addition to considering only a single binary covariate, our simulation study has other limitations. Namely, the values of $TPF_{z|x}$ and $FPF_{z|x}$ were designed to be in a specific relationship to each other. Also we have not investigated the relationship of the proposed approach with the known indicators of inaccuracy of the estimated effect of exposure (e.g. large point or variance estimates).

The future work in this direction may include more general investigation involving multiple continuous and categorical covariates and eliminating the above mentioned deficiencies. Furthermore, since as many other measures of classification accuracy, the AUC computed from the probabilities predicted by the model overestimates the true AUC [16]. Hence the observed ("apparent") estimate of AUC might provide a poor estimation of the true AUC although the overestimation is mostly evident for the smaller values of AUC which are of less concern. In cases when the knowledge of the true underlying AUC is of interest the standard adjustment techniques, such as for instance cross-validation, can be used for obtaining a more accurate estimator.

# 5.0 CONCLUSION

The AUC of the propensity score model for exposure provides a single, relatively easy to compute, and suitable for various kind of data statistic, which can be used as an important indicator of the accuracy of the estimated effect of exposure on the outcome of interest. The public health importance is that it can be considered as an alternative to the previously suggested (Rubin, 2001) simultaneous consideration of the conditions of closeness of means and variances of the propensity scores in the different exposure groups. Our simulations indicate that the estimated effect of exposure is highly unreliable if AUC of the propensity model is larger than 0.8; at the same time AUCs of less than 0.7 are not associated with any substantial increase of inaccuracy of the estimated effect of exposure.

# BIBLIOGRAPHY

1 Walker AM. Confounding by indication.see comment. Epidemiology. 1996;7(4):335-6.

2 Rosenbaum PR. Observational studies. New York, NY: Springer-Verlag 1995:1-12.

3 Schaefer R. Alternate estimators in logistic regression when the data are collinear. Journal of Statistical Computation and Simulation. 1986(25):75-91.

4 Barker L, Brown C. Logistic regression when binary predictor variables are highly correlated. Statistics in Medicine. 200120(9-10):1431-42.

5 Belsley DA, Kuh E, Welsch RE. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity: Wiley, Inc., New York 1980.

6 Hosmer DW, Lemeshow S. Applied Logistic Regression. 2 ed: Wiley-Interscience Publication 2000.

7 Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983(70):41-55.

8 D'Agostino RB. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998 1998(17):2265-81.

9 Rosenbaum PR, Rubin DB. Reducing bias in observational studies. J Am Stat Assoc. 1984(79):516-24.

10 Connors AFJ, Speroff T, Dawson NV, et al. The effectiveness of right heart catheterization in the initial care of critically ill patients. JAMA. 1996(276):889-97.

11 Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. Am J Epidemiol. 1999(150):327-33.

12 Rubin DB. Using Propensity Scores to Help Design Observational Studies: Application to the Tobacco Litigation. Health Services & Outcomes Research Methodology. 2001(2):169-88.

13 Rosenbaum PR, Rubin DB. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. J R Stat Soc. 1983(45):212-8.

14 Zhou X, Obuchowski NA, McClish DK. Statistical Methods in Diagnostic Medicine. New York: John Wiley & Sons, Inc., 2002.

15 Weitzen A, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf. 2004(12):841-53.

16 Yousef, W.A., Wagner, R.F., Loew, M.H. (2005). Estimating the uncertainty in the estimated mean area under the ROC curve of a classifier. Pattern Recongnition Letters, 26, 2600-2610

17 Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996 Dec;49(12):1373-9.