Controlling for Confounding Via Propensity Score Methods Can Result in Biased Estimation of the Conditional AUC: A Simulation Study

Hadiza I. Galadima
Old Dominion University, hgaladim@odu.edu

Donna K. McClish
Virginia Commonwealth University, donna.mcclish@vcuhealth.org

Original Publication Citation

Galadima, HI, McClish, DK. Controlling for confounding via propensity score methods can result in biased estimation of the conditional AUC: A simulation study. *Pharmaceutical Statistics*. 2019; 18: 568– 582. https://doi.org/10.1002/pst.1948

Received: 12 October 2017 | Revised: 8 March 2019 | Accepted: 12 March 2019
DOI: 10.1002/pst.1948

MAIN PAPER

WILEY

Controlling for confounding via propensity score methods can result in biased estimation of the conditional AUC: A simulation study

Hadiza I. Galadima[1] | Donna K. McClish[2]

[1] School of Community and Environmental Health, College of Health Sciences, Old Dominion University, Norfolk, Virginia

[2] Department of Biostatistics, Virginia Commonwealth University, Richmond, Virginia

Correspondence
Hadiza I. Galadima, School of Community and Environmental Health, Old Dominion University, 4608 Hampton Blvd., Health Sciences Bldg., Norfolk, VA 23529-0500.
E-mail: hgaladim@odu.edu

Funding information
NHLBI, Grant/Award Number: 1R01HL64122-01A1

SUMMARY

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 random in observational studies, comparisons of outcomes between exposed and nonexposed 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 conditional odds ratio and hazard ratio. However, research is lacking on the performance of propensity score methods for covariate adjustment when estimating the area under the ROC curve (AUC). In this paper, AUC is proposed as measure of effect when outcomes are continuous. The AUC is interpreted as the probability that a randomly selected nonexposed subject has a better response than a randomly selected exposed subject. A series of simulations has been conducted to examine the performance of propensity score methods when association between exposure and outcomes is quantified by AUC; this includes determining the optimal choice of variables for the propensity score models. Additionally, the propensity score approach is compared with that of the conventional regression approach to adjust for covariates with the AUC. The choice of the best estimator depends on bias, relative bias, and root mean squared error. Finally, an example looking at the relationship of depression/anxiety and pain intensity in people with sickle cell disease is used to illustrate the estimation of the adjusted AUC using the proposed approaches.

KEYWORDS
AUC, bias, covariate adjustment, propensity score, PS variable selection

1 | INTRODUCTION

Evaluating association between exposure and outcome in observational studies is growing rapidly in the health care research field. Because assignments to exposure are not random in observational studies, any comparisons of outcomes between exposed and nonexposed subjects must account for confounders. 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 must use statistical methods to control for bias and confounding.

When study outcomes are continuous and follow a normal distribution, the mean difference between two populations is a well-known measure of treatment effect. However, there is a growing interest in the medical literature about the use of the probability that a randomly selected participant in a treatment or risk factor exposed group has a better response than a randomly selected participant in the control or nonexposed group. This probability is equivalent to the area under the receiver operating characteristic (ROC) curve and is sometimes denoted by $AUC = P(X > Y)$ where X is the response in the treatment group and Y the response in the nontreated group. Hauck et al[7] introduced the concept of $P(X > Y)$ to assess treatment effects after noting that standard tests may fail to identify important treatment difference. They believe that 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 captured when AUC is used as a measure of effect instead of the simple difference of means.[7] Acion et al,[1] Kraemer and Kupfer,[2] and McGraw and Wong[3] have also shown that AUC may be clinically more meaningful than the change in means, the latter represents the magnitude of the mean difference but does not tell patients their chance to improve under the new treatment. They variously described AUC as a "measure that presents good qualities of meaning, simplicity, and robustness,"[1] "clinically interpretable and statistically justifiable,"[2] and "a common language effect size statistic."[3] As noted by Tian,[4] there are a few advantages of using $P(X > Y)$ to assess treatment effects over the change in means.[4] First, it is scale-free, making it a reasonable measure of treatment effect no matter how much variability exists between the two populations' responses. Second, she showed that AUC does not change under monotonic transformation. Hence, the theory developed for the original distribution is also valid for transformed distributions. Nunney et al have also shown that the mean difference does not account for variability within the groups being compared.[5] 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.[1,5]

Propensity score (PS) methods have been used for a long time to adjust for confounding variables in order to reduce bias in observational studies.[6,8] As introduced by Rosenbaum and Rubin in 1983, the key property of the PS is to balance observed covariates between two groups in nonrandomized studies so that the groups are comparable in the sense that their baseline covariates have similar distribution; it follows that treatment assignment and observed covariates are conditionally independent, given the PS.

Some common ways of using PSs to reduce confounding are stratification on the PS, matching on the PS, inverse probability weighting using the PS, and covariate adjustment on the PS.[9] In his seminal works on PSs, Peter Austin assessed the performance of these PSs methods to adjust estimates of relative risk, odds ratio, hazard ratio, marginal odds ratio, marginal hazard risk, and difference in means.[9-14] He also investigated the choice of variables to include in the PS models. However, there is no mention in his work or other literature of the performance of the PS when association is quantified by the area under the ROC curve; neither has anyone looked at the appropriate choice of variables in the context of the AUC as an effect size measure.

The objective of this paper, then, is to investigate the performance of the PS methodology to control for confounding when association between exposure and outcomes is quantified by AUC. Additionally, the optimal choice of variables to include in the PS model is examined. Finally, the performance of the PS approach to control for confounding is compared with that of a conventional regression approach to adjust the AUC for confounding.

This research is organized as follows. Section 2 contains definition and notation of AUC and methods for estimating AUC, an overview of PS methods, and AUC regression analysis. Additionally, the performance of the PS methodology is investigated through a simulation study in Section 2. In Section 3, the results of the simulation study are presented. In Section 5, the proposed approaches are applied to data concerning depression and pain in sickle cell disease. Section 4 is a concluding section which summarizes the results of the simulations studies and addresses limitations.

2 | METHODS

2.1 | AUC as a measure of risk effect

In methods for ROC analysis, the AUC is usually used as an index to summarize how well a diagnostic test can discriminate diseased and nondiseased populations.[15,16] AUC is also equivalent to the probability that for a randomly selected pair of diseased and nondiseased individuals, the diagnostic test value is higher for the person with disease.[17,18] More generally, instead of groups or people with disease, discrimination could also be between risk and nonrisk populations, treatment or nontreatment groups, or some other binary indicator of a clinical state.

AUC has been used in two studies for assessing treatments effects of atomoxetine versus placebo in adults with ADHD in a randomized controlled trial.[19,20] In both studies, the authors found that atomoxetine was efficacious in reducing ADHD symptoms (AUC ≈ 0.6, $P < .001$). The AUC value of 0.6 indicates that the probability that a randomly selected subject treated with atomoxetine showed a reduction in ADHD symptoms as compared with a randomly selected placebo-treated subject is approximately 0.6.

This research is restricted to the comparison of two groups: one of subjects with the treatment/risk factor and the other of subjects without the treatment/risk factor. The measure of effect we suggest is $AUC = P(Y_{RF} > Y_{NRF})$ where Y_{RF} and $Y_{NRF}$are continuous responses from a risk-group and a nonrisk group, respectively. The AUC is interpreted as the probability that a randomly selected participant in the risk group has a more extreme response than a randomly selected participant in the nonrisk group. We assume without loss of generality that larger response values are associated with the risk population, and smaller values with the nonrisk population.

2.2 | Methods for estimating AUC

2.2.1 | The unadjusted AUC

A nonparametric estimate of the unadjusted AUC is computed based on the fact that the AUC is equivalent to the two-sample Mann-Whitney U statistic.[15-17,21,22] Let Y_{RF_i} $(i = 1, \dots, n)$ and Y_{NRF_j}, $(j = 1, \dots, m)$ represent two continuous responses from random variables $Y_{RF}$and Y_{NRF}representing n and m subjects in the risk group and the nonrisk group, respectively. The Mann-Whitney U statistic is defined by

$$U = AUC^{unadj} = \sum_{i=1}^{m} \sum_{j=1}^{n} I\left(Y_{RF_i} > Y_{NRF_j}\right)/mn \tag{1}$$

where $I\left(Y_{RF_i} > Y_{NRF_j}\right)$is an indicator function of the number of pairs in which $Y_{RF_i} > Y_{NRF_j}$. More specifically, $I\left(Y_{RF_i} > Y_{NRF_j}\right) = 1$ *if* $Y_{RF_i} > Y_{NRF_j}$, and 0 otherwise. The variance of the unadjusted AUC is calculated based on a formula suggested by Delong et al for estimating variances of AUCs based on the properties of the Mann-Whitney statistic.[23]

2.2.2 | AUC controlling for covariates

Brumback et al[22] defined the adjusted treatment effect in a clinical trial as $AUC^{adj} = P(Y_A > Y_P | X_A = X_P = X)$ where subscripts A and P represent treatment arm, T = A (active) or P (placebo). If the result depends on X, there is effect modification, else not, in which case AUC^{adj} will be a constant. They note that the treatment effect, if measured as the expected differences in means, should be equal regardless of whether covariates are taken into account, at least in randomized trials, because E(X|T = A) = E(X|T = P). But this may not be the case when the AUC is the effect measure, because the variance of the estimated difference may be decreased with adjustment. In the context of observational studies and risk factors, where adjusted AUC could be written$AUC^{adj} = P(Y_{RF} > Y_{NRF} | X_{RF} = X_{NRF} = X)$, while E(X|RF) = E(X|NRF) will most often not be true in observational studies, it should be true if also conditioning on the PS. Nevertheless, because of the variance, the marginal AUC will not equal the adjusted AUC. In essence, the AUC as an effect measure appears to not be collapsible, and not adjusting for covariates associated with the outcome will lead to attenuation of the AUC, ie, values closer to the null (0.5). This result is consistent with results regarding other nonlinear models such as logistic and proportional hazards regression, particularly when not all covariates associated with Y (outcome) are included.[9,14,24,25]

Brumback et al developed an approach to control for covariates when the covariate of interest, X, is discrete; using this, they developed the following approach:

1) Each level of the discrete covariate $X$is considered as a stratum s where $s = 1, \dots, S$, and Srepresents the total number of strata;

2) Within each stratum*s*, 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 for all subjects i,j in the stratum;

3) The adjusted AUC is the sum of all the indicator functions of the the number of pairs with higher values for the risk factor group, ie, $I\left(Y_{RF_i} > Y_{NRF_j}\right)$within each strata divided by the sum of the product of the number of subjects in the risk factor group and nonrisk 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 $N = \sum_{s=1}^{S} m^s n^s$and m^s and n^s are the number of subjects in the risk factor and nonrisk factor group in stratum s, respectively.[22]

Janes et al proposed a covariate-adjusted measure of classification accuracy.[26] Their approach in estimating the AUC controlling for confounding is based on the concept of placement values (PVs). Let $Y_{R\ F}$ and $Y_{N\ R\ F}$ be two continuous normal responses arising from a risk factor population and a nonrisk factor population, respectively. The 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 Zdenote a vector of covariates for each subject. 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 the covariates$\boldsymbol{Z}$. This is done by specifying a linear model $Y_{NRF} = \beta_0 + \boldsymbol{Z\beta_1} + \varepsilon$ where 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, the PVs($PV_{RF,Z)}$ are computed. The PV is the standard normal CDF of $\left(Y_{RF} - \widehat{\beta}_0 - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\right)/\widehat{\sigma}$; hence, $\widehat{PV}_{RF,Z} = \Phi\left\{\left(Y_{RF} - \widehat{\beta}_0 - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{\boldsymbol{1}}\right)/\widehat{\sigma}\right\}$ where$\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\sigma}$ are the regression coefficient 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 PVs:$AUC = \sum_{i=1}^{n_{RF}} \widehat{PV}_{RF,Z}/n_{RF}$where n_{RF} is the number of case observations. This approach requires the assumption that the outcome and the covariates be linearly related.

In this research, our approach in estimating the adjusted AUC is based on Janes et al's method but in the context of epidemiologic research where the risk effect is quantified by $AUC = P(Y_{RF} > Y_{NRF})$.

2.3 | Overview of propensity score methods

The concept of PSs was introduced by Rosenbaum and Rubin, as a tool to reduce bias in observational studies.[6,8] It is a tool that balances observed covariates between two groups in order to create the same probability structures as that achieved by a "randomized" experiment. The PS is defined as the conditional probability of assignment to a particular group given a vector of observed covariates. Suppose each subject in the cohort has a vector of observed covariates $\boldsymbol{Z}$, and an indicator of risk status Tsuch that $T = 1$ if subject has the risk factor and $T = 0$ if subject does not have the risk factor. Then, the PS, $e(x) = \Pr(T = 1|\boldsymbol{Z})$, is the probability that a subject with covariates $\boldsymbol{Z}$ is in the risk factor group. [6] While several approaches exist to estimate a PS, logistic regression is the method used most often. Logistic regression models the probability of having the risk factor as a function of a set of the observed covariates $\boldsymbol{Z}$. The PS is then computed as the expected probability of being in the risk group, conditional on $\boldsymbol{Z}$.

Once the PS has been estimated, it is used as a variable in an analysis to control for confounding when estimating risk effect. Common PS analysis methods include stratification, matching, and covariate adjustment on the PS, which are the methods considered here.

When using PSs to control for confounding, where the treatment effect is measured with means, odds ratios, relative risk, or hazard ratios, estimands focus on the average treatment effect (ATE) or the average treatment effect for the treated (ATT). In particular, for those measures, one-to-one matching on PS allows one to estimate ATT, while stratification allows one to estimate either ATE, when equal weights of 1/k for the k strata are used to combine results, or ATT if weights are equal to the proportion of treated/at risk subjects in each stratum.[9,27] Stratum weights used here with AUC are functions of the number of subjects in each strata (see formula *(2)* below). In general, whether ATT or ATE should be the desired estimand depends on the research context. In the context of the AUC measure of treatment effect, comparable concepts more likely to apply may be better thought of as AUC treatment effect for the population (ATE) or AUC treatment effect for the treated (ATT). [9] In this work, for simplicity, we assume constant treatment effect and explore the consequences of different PS adjustment methods in terms of AUC measures.

2.3.1 | Stratifying on the propensity score

The basic idea of stratifying on the PS is to group subjects into approximately equal-size groups determined by the quintiles of the estimated PS. These groups are considered to be homogeneous as subjects in each group are expected to have similar PSs. 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 when the comparison is a difference of means.[8,28] Once the subjects have been grouped into strata based on their PSs, we used the technique proposed by Brumback et al, as described in Section 2.3.2, to estimate the adjusted AUCs within each quintile, and the five estimated AUCs are pooled into one overall AUC to estimate the treatment effect.[22] The technique of Brumback et al is based on determining strata by the quintiles of the estimated PSs. It can readily be shown that the proposed adjusted AUC as described in Section 2.2.2 is a weighted average of the stratum-specific AUCs, given by

$$AUC^{adj}_{Stratified} = \sum_{s=1}^{S} w_s AUC_S \tag{2}$$

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 nonrisk factor group in stratum s, respectively. $S = (1,2,3,4,5)$correspond to the quintiles of the PS. Thus, we see that the overall estimated risk effect for the outcome is a weighted average of the (five) stratum-specific risk effects.

The variance of the adjusted PS stratified AUC is given by

$$Var\left[AUC^{adj}_{Stratified}\right] = \sum_{s=1}^{S} (w_s)^2 Var[AUC_S]. \tag{3}$$

2.3.2 | Matching on the propensity score

In PS matching, matched pairs of risk factor and nonrisk factor subjects are created such that pairs have similar PSs. This technique is also known as one-to-one matching. The treatment effect, AUC in this case, is then estimated from the resultant matched sample. The adjusted risk effect, AUC, must include the matching variables in analysis.[29] We incorporate the PS in the matched sample based on the method of Janes et al[26] for accommodating covariates in ROC analysis as described in Section 2.3.2. The risk group effect is estimated in the matched sample as the mean of the PVs for each subject with PS $PS = Z$ in the risk group:

$$AUC^{adj}_{matched} = \sum_{i=1}^{n_{RF}} \widehat{P}V_{RF,Z}/n_{RF} \tag{4}$$

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 PS in the risk group are given by $\widehat{P}V_{RF,PS} = \Phi\left\{\left(Y_{RF} - \widehat{\beta}_0 - \widehat{\beta}_1 PS\right)/\widehat{\sigma}\right\}.\widehat{\beta}_0, \widehat{\beta}_1$ *and* $\widehat{\sigma}$are the estimates of regression coefficients and the root mean squared error, respectively, from the observations in the nonrisk group. These estimates were obtained through a regression model of the response Y_{NRF} in the nonrisk group as a function of the PS. The model is given by$Y_{NRF} = \beta_0 + \beta_1 PS + \varepsilon$, where$\varepsilon \sim N(0, \sigma^2)$. The variance estimates of the adjusted AUC were obtained via bootstrapping using 1000 bootstrap samples of the original observations.

2.3.3 | Covariate adjustment using the propensity score

In the PS covariate adjustment method, the outcome is regressed on two independent variables: an indicator variable Tdenoting the risk status group and the estimated PS. The estimated risk effect is obtained from the regression coefficient for risk status.

The risk group effect is estimated by regressing the outcome variable on the estimated PS and the variable representing risk group status T using the regression method developed by Janes et al described in Section 2.3.2: $AUC^{adj}_{CovAdjust} = \sum_{i=1}^{n_{RF}} \widehat{P}V_{RF,Z}/n_{RF}$.[26] The standard errors for the estimated AUC were obtained by bootstrapping the data.

2.4 | AUC regression adjustment

Rather than summarizing covariates with a PS, direct adjustment for the individual covariates of interest can also be used to estimate the adjusted AUC. Specifically, the outcome is modeled as a function of an indicator variable denoting the risk group status and a set of independent covariates, again using the method of Janes et al.[26] We refer to this method as the "direct AUC regression adjustment" method.

2.5 | Design of simulation study

To examine the performance of different PS methods and models for estimating conditional treatment or risk effects, data were simulated using a framework similar to those used by Austin et al.[10,14] Data were 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 Table 1.

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 were related "strongly" or "moderately" 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,$c_6$were "strongly" or "moderately" related to the outcome. The eight variables b_1,c_1,b_2,c_2,b_4,c_4, and b_5,$c_5$were related to both risk group and the outcome and were 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.[29,30] 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,[29] and 1.5 and 1.25 for continuous covariates.[10]

Similarly, the association between outcome and a binary variable was measured with the point-biserial correlation; the association between outcome and a continuous variable is measured with 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 proposed these guidelines for interpreting the magnitude of correlation coefficients.[31]

In summary, 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 ratio values of 2, 1.5, and 1 for a strong, moderate, and no association between a given covariate and the risk factor group.

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

PS-Model 1:. This model included 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 included 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.

TABLE 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

PS-Model 3:. This model included all eight 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 included all 18 generated variables: $b_1 - b_9$ and $c_1 - c_9$.

Step 2:. A risk factor status *T*was generated for each subject. Data were simulated such that the logit of the probability of having the risk factor for the ith 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 (P_{group}) was related to the 12 baseline covariates that are strongly and moderately associated with the risk group, ie, $(b_1, b_2, b_4, b_5, b_7, b_8, c_1, c_2, c_4, c_5, c_7, c_8)$through the logit model:

$$logit = \log\left(\frac{P_{group}}{1-P_{group}}\right) = \beta_0 + \beta_1 b_1 + \beta_2 b_2 + \beta_4 b_4 + \beta_5 b_5 + \beta_7 b_7 + \beta_8 b_8 + \alpha_1 c_1 + \alpha_2 c_2 + \alpha_4 c_4 + \alpha_5 c_5 + \alpha_7 c_7 + \alpha_8 c_8. \tag{5}$$

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

$$P_{group} = \frac{\exp(logit)}{1 + \exp(logit)}. \tag{6}$$

The risk factor status *T*for each of the N subjects was generated from a Bernoulli distribution with a parameter (P_{group}), ie, *T*~*Bernouilli* (P_{group}). The risk factor status vector is computed by comparing the estimated probability of group assignment (P_{group}) to a random variable U generated from*Uniform*(0,1). We assign $T = \begin{cases} 1 & \text{if } U \le P_{group} \\ 0 & \text{if } U > P_{group} \end{cases}$.

Step 3:. For each of the N subjects, a continuous outcome Y conditional on risk factor status *T*was generated through the following linear model:

$$Y = \alpha_0 + \delta T + \beta_1^* b_1 + \beta_2^* b_2 + \beta_3^* b_3 + \beta_4^* b_4 + \beta_5^* b_5 + \beta_6^* b_6 + \alpha_1^* c_1 + \alpha_2^* c_2 + \alpha_3^* c_3 + \alpha_4^* c_4 + \alpha_5^* c_5 + \alpha_6^* c_6 + \varepsilon. \tag{7}$$

Each regression coefficient was estimated assuming the outcome Y and the single covariate X (ie, $b_1 - b_9$,$c_1 - c_9$) were related through a regression equation. The derived formula to estimate theβ^**s*is given by $\beta^* = \rho\frac{\sigma_\varepsilon}{\sigma_x}\sqrt{\frac{1}{(1-\rho^2)}}$ where ρ is the Pearson product-moment correlation between a covariate x and the outcome Y, and σ_x and σ_ε are the standard deviations of the covariate of interest and the error term, respectively. The effect on outcome of the risk group compared with the nonrisk group is quantified by AUC statistic through δTin Equation (7). Hence, the effect size is given by $\delta = \sigma_\varepsilon\sqrt{2}\Phi^{-1}(AUC_0)$that is $\delta$is a function of the true AUC which is denotedAUC_0, and Φ denotes the standard normal cumulative distribution function.

2.6 | Simulating data

A sample of size N = 500 was considered in this simulation study; for each of the N subjects, we randomly generated: (a) 18 independent baseline covariates such that nine of them are dichotomous variables from a Bernoulli distribution with parameter 0.5: $(b_1, b_2, b_4, b_5, b_7, b_8, b_9)$~*Bernouilli*(0.5)and the other nine are continuous from a standard normal distribution: $(c_1, c_2, c_4, c_5, c_7, c_8, c_9)$~*N*(0,1). Each of the 18 covariates varied in their association with the risk group and the outcome as described in Table 1. (b) A risk factor status for each of the N subjects was generated using Equations (5) and (6) as described in Section 2.6.β_0 was 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 simulations. 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. (c) Finally, for each subject N, a continuous outcome conditional on the risk factor status T was generated as in Equation 7 such that

$Y = \delta T{+}4.6b_1{+}4.6b_2{+}4.6b_3{+}2.6b_4{+}2.6b_5{+}2.6b_6{+}2.3c_1{+}2.3c_2{+}2.3c_3{+}1.3c_4{+}1.3c_5{+}1.3c_6{+}\varepsilon \text{where} \varepsilon{\sim}N(0{,}4)$. The error variance 4 for the outcome model has been chosen after an iterative process to induce the desired parameters estimates values. The data generating process described here was repeated 2500 times. All data generation and analyses were completed using SAS version 9.4. As evaluation criteria for the performance of the estimated AUC, we considered relative bias and root mean squared error (RMSE) across 2500 simulated data sets (replications).

3 | RESULTS

Results of the simulation study are given in Table 2. This table displays the AUC estimates, relative bias, and RMSE for three true values of AUC, four adjustment methods, and four covariate set models. The AUC estimates are the estimated mean risk effect across the 2500 simulated data sets for each propensity score method and for each model. The obtained crude estimates when AUC is 0.5, 0.7, and 0.9 are 0.63, 0.73, and 0.85, respectively. They are biased positively when the true risk group effect are 0.5 and 0.7 but biased negatively when the true AUC is 0.9.

When stratification on the quintiles of the PS is used, we observe four things: (a) the amount of bias is somewhat similar for each true effect regardless of the PS model used; (b) the risk effect is overestimated when there is no effect

TABLE 2 Results of the simulation study: AUC estimates, relative bias, and RMSE for AUC = 0.5, 0.7, and 0.9

		AUC = 0.5			AUC = 0.7			AUC = 0.9		
Methods	Models[a]	AUC Estimates	Relative Bias	RMSE	AUC Estimates	Relative Bias	RMSE	AUC Estimates	Relative Bias	RMSE
Unadjusted		0.63	25.85	0.13	0.73	4.31	0.04	0.85	5.91	0.06
Stratification										
	Model 1	0.55	10.04	0.08	0.64	−8.95	0.09	0.79	−12.23	0.12
	Model 2	0.55	9.77	0.08	0.65	−7.24	0.09	0.81	−10.25	0.10
	Model 3	0.55	10.23	0.08	0.65	−6.98	0.08	0.81	−9.99	0.10
	Model 4	0.55	9.90	0.08	0.64	−9.00	0.08	0.79	−12.33	0.12
Matching										
	Model 1	0.50	−0.67	0.02	0.61	−12.33	0.09	0.76	−15.14	0.14
	Model 2	0.50	0.83	0.02	0.63	−9.80	0.07	0.79	−12.25	0.11
	Model 3	0.49	−1.97	0.02	0.62	−11.04	0.08	0.79	−12.43	0.11
	Model 4	0.50	−0.29	0.02	0.62	−11.99	0.09	0.77	−14.86	0.13
PS covariate adjustment										
	Model 1	0.51	1.34	0.02	0.63	−10.65	0.08	0.78	−13.82	0.13
	Model 2	0.50	−0.60	0.02	0.62	−10.72	0.08	0.79	−12.78	0.12
	Model 3	0.48	−3.54	0.03	0.61	−12.19	0.09	0.78	−13.19	0.12
	Model 4	0.51	2.34	0.02	0.63	−10.13	0.07	0.78	−13.63	0.12
Regression adjustment										
	Model 1	0.51	1.40	0.02	0.65	−6.71	0.05	0.83	−8.26	0.08
	Model 2	0.50	0.26	0.02	0.70	0.26	0.02	0.90	0.15	0.01
	Model 3	0.49	−2.76	0.02	0.63	−9.41	0.07	0.81	−9.69	0.09
	Model 4	0.52	3.33	0.03	0.71	1.99	0.02	0.91	0.73	0.01

[a] Model 1 includes all variables associated with risk factor; Model 2 includes all variables associated with outcome; Model 3 includes variables associated with both risk factor and outcomes; Model 4 includes all variables.

(True AUC = 0.5) and underestimated when the true effects are 0.7 and 0.9; (c) the risk estimate when truth is 0.7 is associated with the least bias; and (d) bias was somewhat less when using matching or PS covariate adjustment.

When matching on the PS is used, when there is no effect (AUC = 0.5), the bias is almost null, but it is not the case when the true AUC was 0.7 or 0.9, where relative bias is generally 10% to 15%. Models 2 and 3 appear to have the least bias when the effect is non-null.

When covariate adjustment on the PS is used, the findings are similar to the previous ones. When AUC is 0.5, the results are similar to those found with matching. However, PS model 2 or 4, ie, models including variables associated with outcome seem to have the least bias in the non-null case.

From these results, it appears that stratifying, matching, and covariate adjustment on the PS resulted in biased estimation of AUC when true effects were non-null. 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%.

Finally, for comparative purposes, we investigated risk effects estimated from directly regressing individual covariates. The mean estimated risk effects perform better than those estimated from the PS methods. 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 area 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, as seen in Table 2, these models had increased RMSE when true effects were 0.7 and 0.9.

4 | Case Study

4.1 | Data sources

Sickle cell disease (SCD) is a genetic disease that manifests itself in pain, both acute and chronic. Patients with sickle cell disease are more likely than the general population to have depressive or anxiety disorders.[32] Because chronic pain has been shown to be associated with depression, it is of interest to know whether SCD patients with depression/anxiety have more intense pain than those with SCD but without depression.

The Pain in Sickle Cell Epidemiology Study (PiSCES) is a longitudinal study of pain in sickle cell disease. The methods of PiSCES have been described in detail elsewhere.[33] Briefly, subjects were enrolled in both medical center and community settings between July 2002 and August 2004, with most from the Richmond and Tidewater areas. Subjects aged 16 years and older were eligible for enrollment. A total of 308 people with SCD initially were enrolled. Baseline information including demographics and psychosocial measures was collected, and daily pain diary data were completed for up to 6 months. Covariates considered for analysis included demographics (age, sex, marital status, income, education, and genotype), SF-36 summaries of physical and mental health related quality of life (HRQOL—values range from 0 to 100 with higher being better), SCD stress, (ranging from 0 to 40, high values indicating more stress), SCD coping (values are mean of subscales and range from 0 to 6), number of somatic symptoms ranged from 0 to 18, while number of SCD comorbidities could be as high as 19, and ridicule, a measure of negative social exchange from the Test of Negative Social Exchange measure ranges from 0 to 4. Among other items, the diary asked subjects to report the worst sickle cell pain intensity experienced during the previous 24 hours on a scale from 0 (no pain) to 9 (unbearable).

A total of 232 subjects met the inclusion criteria of filling at out at least 1 month of diaries, 220 of these reported at least 1 day with pain score greater than zero. After removal of subjects due to missing covariates, there were N = 196 subjects for analysis. The study was approved by the Institutional Review Board of Virginia Commonwealth University.

4.2 | Statistical analyses

For each subject, PSs were estimated by fitting a logistic regression to predict depression/anxiety status, as a function of 16 baseline covariates. We constructed four different PS models which included different combinations of measured covariates: PS model 1 (PS-M1) included variables associated with the depression/anxiety status group. The association between the covariates and the depression/anxiety status was 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, mean pain intensity during pain days over a 1 to 6-month period. The association between mean pain intensity and the continuous covariates was measured using a Pearson correlation, and a t-test was used to test association between the outcome and the categorical covariates. PS model 3 (PS-M3) included variables associated with both depression/anxiety status and mean pain intensity, ie, all common covariates to the previous two models. PS model 4 (PS-M4) included all measured variables.

To compare patients with SCD who have depression/anxiety with patients who do not, all four methods described earlier were used to estimate the adjusted probability that patients with depression/anxiety would have higher average intensity of pain over a 1 to 6-month period compared with those without depression$P(Y_{Depr} > Y_{NoDepr}|X)$. First, subjects were stratified based on the quintiles of the PS, and the adjusted AUC was computed as described in Section 2.4. Second, we estimated the adjusted depression/anxiety effect via AUC in the PS matched sample as described in Section 2.4 as well. Third, the risk group effect was estimated under the covariate adjustment on the PS method using the method described in Section 2.3.3. Finally, for comparison purpose, the direct AUC regression method was used to adjust for covariates in directly modeling covariates effects on the response as described in Section 2.4. For this method, we considered four separate regression models as well. Those models were similar to the four PS models described earlier.

4.3 | Results

The summary statistics of the baseline covariates between subjects with depression and those without depression are presented in Table 3. The descriptive analysis reveals that patients with depression are older with worse physical and mental HRQOL scores, more somatic complaints, higher active, passive, and affective coping scores, yet lower income ($P < .05$). Pain intensity on pain days was associated with physical and mental HRQOL, somatic complaints, SCD comorbidities, and income ($P < .05$). There was no statistically significant relationship between depression and sex, marital status, education, genotype, number of comorbidities, stress, social support, and ridicule.

As seen in Table 4, PS-M1 contained nine covariates, PS-M2 had nine variables, and seven variables for PS-M3 and PS-M4 contained all 16 variables. Logistic regression was used to estimate the PSs. The crude AUC between depression groups was 0.6016 with a 95% confidence interval of (0.5128-0.6905). Because the confidence interval does not contain the null value 0.5, we might conclude that when not adjusting for any other variables, for two randomly chosen patients one with depression and the other not, the probability is 0.6016 that the mean pain intensity from the SCD patient with depression is higher than the mean pain intensity for patient without depression; that represents a statistically significant chance of increased pain intensity for those with depression at baseline.

The adjusted estimates using the four different methods are reported in Table 5. Using stratification on the quintiles of the PS, the adjusted estimates of $P(Y_{NS} > Y_S|X)$ range from 0.5666 to 0.66634 for different PS models. In contrast to the unadjusted AUC, all four 95% 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, ie, we fail to reject$H_o : AUC = 0.5$. Propensity score matching resulted in the formation of 33, 39, 41, and 32 pairs of subjects out of a possible maximum of 60 for PS models 1, 2, 3, and 4, respectively. The range of adjusted estimates was lower than with stratification for each model, ranging from 0.4625 to 0.5180. Nevertheless, the results were consistent between matching and stratification in that all confidence intervals contain the null value (0.5). Using covariate adjustment on the PS, results were similar to what we found with matching. We also evaluated the use of AUC regression to directly model the covariates on the response. AUC results ranged from 0.5031 to 0.5554. All four confidence intervals are consistent in containing the null value of 0.5. Because, in this analysis, the null appears to be true, it is not surprising that all four methods of adjustment are in agreement, that adjusting for baseline covariates, SCD patients with sickle cell who are depressed do not report significantly different levels of pain intensity.

5 | DISCUSSION

The primary objective of this research was to evaluate the performance of PS methods to control for confounding when estimating the area under the ROC curve. The simulation study demonstrated that when AUC is used as measure of risk factor effect, conditioning on the propensity often results in biased estimation of the true conditional risk factor effects. When the true effect was null, ie, AUC was 0.5, matching on the PS and covariate adjustment on the PS were associated with little or no bias; slightly more bias was incurred when using the method of stratifying on the PS. When the true effect was different from the null effect, the estimated AUC were all associated with large bias for all different methods and models.

TABLE 3 Baseline characteristics of the study sample by depression/anxiety status, and relationship of characteristics with pain intensity

	Depression/Anxiety N = 60 (30.6%)	No Depression/Anxiety N = 136 (69.4%)	P-Value w Depr/Anxiety	Correlation w Pain	P-Value w Pain
Demographic variables					
Age (years)	38.2 ± 12.1	32.6 ± 10.1	0.0008	−0.011	0.8785
Sex					
Male	18 (30.0%)	56 (41.2%)	0.1369	--	0.8173
Female	42 (70.0%)	80 (58.8%)			
Marital status					
Married	10 (16.7%)	34 (25.0%)	0.1975	--	0.3638
Not married	50 (83.3%)	102 (75.0%)			
Education					
<HS	8 (13.3%)	17 (12.0%)	0.5726	--	0.8483
HS	25 (41.7%)	47 (34.6%)			
>HS	27 (45.0%)	72 (52.9%)			
Income					
<$10 000	34 (56.7%)	40 (29.4%)	0.0008	--	0.0008
$10 000-20 000	14 (23.3%)	32 (23.5%)			
$20 000-30 000	3 (5.0%)	27 (19.8%)			
>$30 0000	9 (15.0%)	37 (27.2%)			
Psychosocial variables					
Physical HRQOL (PCS)	31.4 ± 9.4	36.8 ± 9.8	0.0004	−0.3439	<0.0001
Mental HRQOL (MCS)	39.5 ± 11.7	50.6 ± 9.0	<0.0001	−0.1517	0.0337
Active coping	3.2 ± 1.1	2.8 ± 1.0	0.0255	0.215	0.0025
Passive coping	4.2 ± 9.0	3.9 ± 1.1	0.0411	0.1701	0.0172
Affective coping	3.1 ± 1.0	2.3 ± 1.2	<0.0001	0.2121	0.0028
Stress	21.7 ± 8.3	18.8 ± 10.2	0.0548	0.0802	0.264
Social support	5.6 ± 1.3	5.6 ± 1.3	0.9016	−0.0413	0.5659
Ridicule	0.5 ± 0.7	0.6 ± 0.8	0.2454	0.1282	0.0734
Disease-specific variables					
Genotype (SS)	41 (68.3%)	105 (77.2%)	0.1891	--	0.3851
No. somatic symptoms	9.6 ± 3.5	6.1 ± 3.4	<0.0001	0.1572	0.0215
Number of comorbidities	2.4 ± 1.8	2.6 ± 1.8	0.4658	0.2253	0.0078

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

Correlation with pain only computed for continuous variables. *P*-values reported are for t-tests/ANOVAS comparing pain intensity for levels of categorical variable.

In a simulation study conducted by Austin et al, they found that controlling for covariates using PS methods when estimating conditional odds ratio and conditional hazard ratio resulted in biased estimation of the true effect.[14] Our simulation, which focuses on "conditional AUC" would likely be expected to find similar results. Thus, our results are not totally unexpected. This study is the first to evaluate the performance of different PS methods for controlling for covariates when estimating area under the ROC curve, ie, $P(Y_{RF} > Y_{NRF}|X)$. Due to the suggestions in the epidemiologic literature to report $P(Y_{RF} > Y_{NRF})$ as a measure of association/treatment effect[18] and to the common practice of using PS methods to control for confounding in observational studies, it is of practical importance that the statistical properties of PS estimators as a means of adjusting AUC estimates be understood.

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

TABLE 4 Selection of variables entering different propensity score models

Covariates	PS Model 1	PS Model 2	PS Model 3	PS Model 4
Age (years)	✓			✓
Physical HRQOL (PCS)	✓	✓	✓	✓
Mental HRQOL (MCS)	✓	✓	✓	✓
Active coping	✓	✓	✓	✓
Passive coping	✓	✓	✓	✓
Affective coping	✓	✓	✓	✓
Somatic symptoms	✓	✓	✓	✓
Number of comorbidities		✓		✓
Stress	✓			✓
Social support				✓
Ridicule		✓		✓
Genotype				✓
Income	✓	✓	✓	✓
Married				✓
Education				✓
Sex				✓

TABLE 5 Effect estimates using different methods and models

Methods	Models[a]	AUC	Standard Error	95%CI
Unadjusted		0.6016	0.0529	0.5128-0.6905
Stratification using PS				
	Model 1	0.6634	0.1137	0.4406-0.8862
	Model 2	0.5572	0.1273	0.3077-0.8067
	Model 3	0.5666	0.1209	0.3444-0.7889
	Model 4	0.6421	0.1599	0.4211-0.8631
Matching using PS				
	Model 1	0.4625	0.0684	0.3281-0.5961
	Model 2	0.5180	0.0651	0.3907-0.6457
	Model 3	0.4911	0.0630	0.3688-0.6157
	Model 4	0.4837	0.0713	0.3450-0.6246
PS covariate adjustment				
	Model 1	0.4445	0.0691	0.3070-0.5778
	Model 2	0.5181	0.0596	0.4038-0.6373
	Model 3	0.4878	0.0574	0.3744-0.5993
	Model 4	0.4863	0.0779	0.3336-0.6397
Regression adjustment				
	Model 1	0.5031	0.0562	0.3926-0.6127
	Model 2	**0.5554**	**0.0551**	**0.4473-0.6635**
	Model 3	0.5237	0.0552	0.4160-0.6323
	Model 4	0.5311	0.0574	0.4173-0.6422

[a] Model 1 includes all variables associated with risk factor; Model 2 includes variables associated with outcome, Model 3 includes variables associated with both risk factor and outcomes; Model 4 includes all variables.

Bold means Model 2 has the best performance.

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 as 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 PS models, Brookhart et al[34] as well as Austin[35] found that a PS model which includes covariates associated with outcome or the true confounders resulted in a larger number of matched samples, a greater precision of the estimated treatment effect, and a lower bias. Furthermore, Austin found that variables associated with treatment exposure but not the outcome increased the MSE of the estimated relative risk.[11] Interestingly, there is some work that suggests that including variables strongly associated with treatment/risk factor but not outcome (these are often referred to in the economic literature as instrumental variables) may result in bias amplification with increases in bias as well as variance of effect measures, including some nonlinear effect measures such as odds ratios.[36] While this has not specifically been examined for the AUC measure, results here suggest that including instrumental variables in PSs and covariate adjustment (such as b7, c7 in model 3) results bias amplification.

A third objective was to compare the performance of the PS approach with that of using the individual covariates of interest to adjust estimates of $P(Y_{RF} > Y_{NRF})$. The results of our simulation study show that the AUC regression models including all covariates associated with outcomes (models 2 and 4) have the best performance and result 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 RMSE when AUC > 0.5. Austin et al advocate that the choice between PS 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.[14] They noted that the conventional regression adjustment estimates conditional treatment effect while the PS estimates marginal treatment effects such as in a randomized trial. As mentioned previously, these only coincide if the effect measure is collapsible, which is not the case for the AUC,[22] nor was it for the odds ratio of hazard ratio.[14]

A limitation to the use of the PS methodology in practice includes the fact that it only controls for observed variables, as with any method. The unobserved variables are accounted for only if they are correlated with the observed covariates. And while the optimal choice of variables for a PS should include those associated with outcome and exclude those that are not (instrumental variables), it may be difficult to determine which variables exhibit these properties. Although the baselines covariates were assumed to be all inclusive and correctly measured in the simulations, this assumption in practice can be more problematic. On the other hand, some PS methods have been shown to be relatively robust to incorrect modeling.

In a systematic review conducted by Weitzen et al covering publications through 2001, using the PS directly as a covariate was shown to be a popular choice of covariate adjustment method.[37] Austin and others in more recent work also consider covariate adjustment as an option for PS adjustment. In fact, Austin has developed methods to assess goodness-of-fit when using the PS as a covariate in order to determine whether the PS model is correctly specified and the baseline characteristic balanced between groups.[38] On the other hand, it should be noted that Schafer and Kang discourage such use because they feel that the interpretability of covariate adjustment in terms of causal inference may be suspect. As they point out, the treatment effect assessed with covariate adjustment is an average response difference between two groups, adjusting for differences in covariates, while causal inference focuses on changes in response when two treatment are applied to the same people. This issue applies equally to using the PS as a covariate, or directly adjusting the AUC with covariates (regression adjustment). While under some circumstances the results are the same, conceptual differences might suggest that this method be avoided.[39] Additionally, an issue with PS covariate adjustment as compared with other PS methods is that the method relies on modeling a relationship of the PS to the outcome (separate from modeling relationship of covariates to treatment/risk factors). Model misspecification (eg, linear or nonlinear) at this level can have serious consequences.[9,39]

Limitations to the simulations in this research include the use of a limited number of AUC values, a single sample size, only one prevalence of risk factor (50%), and equal standard deviation between those with and without the risk factor. Also, for simplicity, only independent variables were considered in the simulation; we could have considered correlated variables as well. In addition, the simulation only considered linear relationships between covariates and outcomes. This is unlikely to be true in real life. We also assumed that the outcome of interest was normally distributed. If the outcome variable is not normally distributed, and cannot be so transformed, other methods would need to be used. We did not assess the amount of overlap of our samples, nor did we assess balance actually achieved. Finally, and perhaps most importantly, the simulation generated data appropriate for conditional treatment effects, not marginal effects. Further research considering marginal effects are likely to find that PS methods will be useful, having much less bias, although some of Austin's work found some bias remained even for estimating marginal odds ratios and hazard ratios.[12,13] So it is not clear what would be found when considering a simulation for looking at marginal effects assessed with the AUC.

In conclusion, given these findings and based on our simulation study, we do not recommend the use of PS methods to provide adjusted estimates of conditional effects when the AUC is used (ie, $P(Y_{RF} > Y_{NRF}|X)$). If interest is in marginal effects, it may be that PS methods will be the method of choice, but that research has not been done. It appears that direct adjustment with the individual covariates of interest can be used to estimate the conditional, adjusted AUC, at least when outcomes are continuous and follow a normal distribution, and covariates exhibit linear relationships with outcomes. Furthermore, AUC regression modeling adjusting for covariates related to the outcome and the model adjusting for all variables lead to unbiased estimation of conditional AUC under these circumstances. But with conventional regression adjustment, one cannot easily determine the overlap of the distributions (common support), thus possibly involving unwitting extrapolation. Results may be sensitive to correct specification of the model, while adjustment with PS models, when appropriate, is more robust.

DECLARATION OF CONFLICTING INTEREST

The authors declare that there is no conflict of interest.

ACKNOWLEDGEMENTS

The authors would like to thank the anonymous reviewers for their insightful comments and recommendations, which have substantially improved this manuscript. This research was supported by a grant from NHLBI: 1R01HL64122-01A1.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Hadiza I. Galadima https://orcid.org/0000-0003-1588-3929
Donna K. McClish https://orcid.org/0000-0003-0344-6377

REFERENCES

1. Hauck WW, Hyslop T, Anderson S. Generalized treatment effects for clinical trials. *Stat Med*. 2000;19(7):887-899.
2. Acion L, Peterson JJ, Temple S, Arndt S. Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. *Stat Med*. 2006;25(4):591-602.
3. Kraemer HC, Kupfer DJ. Size of treatment effects and their importance to clinical research and practice. *Biol Psychiatry*. 2006;59(11):990-996.
4. McGraw KO, Wong SP. A common language effect size statistic. *Psychol Bull*. 1992;111(2):361-365.
5. Tian L. Confidence intervals for P(Y1>Y2) with normal outcomes in linear models. *Stat Med*. 2008;27(21):4221-4237.
6. Nunney I, Clark A, Shepstone L. Estimating treatment effects in a two-arm parallel trial of a continuous outcome. *Stat Med*. 2013;32(6):941-955.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41-55.
8. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *JASA*. 1984;79(387):516-524.
9. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46(3):399-424.
10. Austin PC. The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies. *Stat Med*. 2010;29(20):2137-2148.
11. Austin PC. The performance of different propensity-score methods for estimating relative risks. *J Clin Epidemiol*. 2008;61(6):537-545.

12. Austin PC. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*. 2007;26(16):3078-3094.
13. Austin PC. The performance of different propensity score methods for estimating marginal hazard ratios. *Stat Med*. 2013;32(16):2837-2849.
14. Austin PC, Grootendorst P, Normand ST, Anderson GM. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: a Monte Carlo study. *Stat Med*. 2007;26(4):754-768.
15. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. New York: Oxford University Press; 2003.
16. Zhou X, Obuchowski NA, McClish DK. *Statistical Methods in Diagnostic Medicine*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2011.
17. Hanley JA. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982;143(1):29-36.
18. Colditz GA, Miller JN, Mosteller F. Measuring gain in the evaluation of medical technology. The probability of a better outcome. *Int J Technol Assess Health Care*. 1988;4(04):637-642.
19. Farone SV, Biederman J, Spencer T, et al. Efficacy of atomoxetine in adult attention-deficit/hyperactivity disorder: a drug-placebo response curve analysis. *Behav Brain Funct*. 2005;1(1):16.
20. Farone SV, Biederman J, Spencer TJ, Wilens TE. The drug-placebo response curve: a new method for assessing drug effects in clinical trials. *J Clin Psychopharmacol*. 2000;20.
21. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals Math Stat*. 1947;18(1):50-60.
22. Brumback LC, Pepe MS, Alonzo TA. Using the ROC curve for gauging treatment effect in clinical trials. *Stat Med*. 2006;25(4):575-590.
23. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837-845.
24. Gail MH, Wieand S, Piantadosi S. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*. 1984;71(3):431-444.
25. Greenland S, Robins J, Pearl J. Confounding and collapsibility in causal inference. *Statistical Science*. 1999;14:29-46.
26. Janes H, Longton G, Pepe M. Accommodating covariates in ROC analysis. *The Stata J*. 2009;9(1):17-39.
27. Imbens GW. Nonparametric estimation of average treatment effects under exogeneity: a review. *Rev Econ Stat*. 2004;86(1):4-29.
28. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*. 1968;24(2):295-313.
29. Pepe MS, Fan J, Seymour CW. Estimating the ROC curve in studies that match controls to cases on covariates. *Acad Radiol*. 2013;20(7):863-873.
30. Monson RR. *Occupational Epidemiology*. 2nd ed. Boca Raton: FL CRC Press; 1990.
31. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum Associates: Hillsdale, N.J; 1988.
32. Levenson JL, McClish DK, Dahman BS, et al. Depression and anxiety in adults with sickle cell anemia: the PiSCES project. *Psychosomatic Med*. 2008;70(2):192-196.
33. Smith WR, Bovbjerg VE, Penberthy LT, et al. Understanding pain and improving management of sickle cell disease: the PiSCES study. *J Natl Med Assoc*. 2005;97(2):183-193.
34. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am J Epidemiol*. 2006;163(12):1149-1156.
35. Austin PC, Grootendorst P, Anderson G. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*. 2007;26(4):734-753.
36. Pearl J. Understanding bias amplification [invited commentary]. *Am J Epidemiol*. 2004;174:1223-1227.
37. Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf*. 2004;13(12):841-853.
38. Austin PC. Goodness of fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiol Drug Saf*. 2008;17(12):1202-1217.
39. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychol Methods*. 2008;13(4):279-313.

How to cite this article: Galadima HI, McClish DK. Controlling for confounding via propensity score methods can result in biased estimation of the conditional AUC: A simulation study. *Pharmaceutical Statistics*. 2019;18:568–582. https://doi.org/10.1002/pst.1948