ORIGINAL ARTICLE

Propensity Score Approaches in Quantifying Effects of Treatment from Observational Data

Alvin Duke R. Sy, RN, MSPH and Abubakar S. Asaad, PhD, MAT, MOS, BSE

Department of Epidemiology and Biostatistics, College of Public Health, University of the Philippines Manila

ABSTRACT

Introduction. Despite the growing popularity of utilizing observational studies for determining associations with public health implications, there is limited literature using them for examining and quantifying the effects of exposures or treatments: The study compared traditional regression with scoring approaches in estimating treatment effects considering the noted limitations in the dataset.

Methods. We conducted a secondary analysis of previously collected retrospective cohort data derived from maternal-neonatal dyads delivered prematurely in a tertiary hospital. Propensity scores (PS) were estimated using logistic and boosting regression. These scores were implemented into matching, stratification, and weighting models. The estimated measures of effect from traditional regression and PS-adjusted models were compared using certain metrics (i.e., the width of CI, SE, AIC, BIC). Sensitivity analysis was also performed.

Results. We included data from 562 patients (123 untreated and 439 treated). Both the estimated scores demonstrated satisfactory fit and reduction in the standardized differences between the groups. However, the logit-estimated scores had better prediction (AUC: 0.71 vs 0.66) and forecasting properties (Brier: 0.15 vs 0.17) than the boosting-estimated scores. All generated statistical models demonstrated a reduction in the occurrence of respiratory morbidity among preterm neonates exposed to a single-dose antenatal corticosteroid (ACS) (ORs ranged from 0.37 to 0.59). The estimated average treatment effects (ATE) and effect among those treated (ATET) from various models suggested a small benefit attributed to the single-dose ACS (ATEs range from -0.09 to -0.41; ATETs range from -0.07 to -0.17).

Conclusion. PS estimated using logistic regression performed better than those estimated using machine learning strategies. The matching model using the said scores demonstrated better fit and parsimony over conventional and propensity-adjusted models. Future studies are recommended to improve the application of these analytic techniques in real-world data.

Keywords: propensity scores, machine learning, logistic regression, treatment effects, observational studies

INTRODUCTION

The identification and quantification of treatment or exposure effects in social and health sciences remain to be an important methodological challenge. Experimental studies have always been accepted as the "gold standard" for determining the effects of exposures (e.g., treatments, programs, risk factors) on the occurrence of outcomes. One can draw models to investigate relationships between manipulable exposures to isolate outcome-exposure associations - with the advantages of temporality, and the randomization on treatment allocation guaranteeing that the sampling design will result in having two groups at the beginning of the study – not influenced by their social, clinical, and demographic characteristics. Furthermore, the possibility of

Paper presented in the 4th Graduate Students' Colloquium organized by the University of the Philippines Manila on October 28, 2021 via Zoom Meetings.

Corresponding author: Alvin Duke R. Sy, RN, MSPH
Department of Epidemiology and Biostatistics
College of Public Health
University of the Philippines Manila
625 Pedro Gil Street, Ermita, Manila 1000, Philippines
Email: arsy3@up.edu.ph

accounting for both known and unknown confounders in the association is very likely to be collected with high quality and precision in experimental than observational studies.[1]

However, such relative ease of estimating effect measures has always been a problem for quasi- and non-experimental studies. The unknown mechanism of selecting participants among studies, which do not involve a random allocation method, leads to a non-probabilistic equivalence between the groups at baseline; or these studies tend to create comparison groups who do not possess similar pre-exposure characteristics. Despite, observational studies having the same intent as clinical trials – the presence of these threats to validity results in bias or less precise estimates than those gathered from studies where manipulation and randomization were part of the design. These factors not only made direct computation of treatment or exposure effects more difficult but also weakened the strength and rigor of these estimated measures of effect.[2]

Currently, there is growing interest in the use of observational studies for assessing exposure effects on outcomes – especially when randomization is difficult, the use of controlled experiments is not feasible, and ethico-moral considerations are imperative.[3] Over time, improvements in the creation of better study designs among observational studies such as increasing the number of measured variables, and other methods to account for and provide better solutions to reduce bias and confounding. Regression has appeared to be the standard approach to analysis but the difficulty in satisfying the necessary assumptions such as adequacy of sample size for rare conditions and exposures; non-linearity of certain associations, large sets of risk factors, or confounding variables to the association; and comparability of the study groups.[4]

A relatively recent method used in addressing the aforementioned problem is the use of propensity score analysis. In this approach, the exposed and unexposed individuals are equated based on the estimated risk of receiving the exposure hence, creating multiple matches or strata where units within a certain propensity score, regardless of their exposure status, have comparable pre-treatment, measured characteristics.[5] Thereby, we can assume that the presence of exposure within a certain increment of the propensity score was random, and yielded relatively unbiased estimates of treatment effects.[6]

Moreover, the potential and advantages of machine learning algorithms in the context of propensity score analysis were noted in the literature.[7] Logistic and probit regression have been the staple methods in estimating propensity scores, but algorithms particularly the generalized boosting methods (GBM) can minimize imbalances by taking into account any non-linear or higher-order association between exposure status and the known covariates, iteratively – and with fewer assumptions that have to be satisfied.[8]

However, a caveat of using propensity scores and machine learning techniques is that these have not been often used clearly and intricately in the analysis of epidemiologic and public health data.[9] This is attributed to the current lack of a strong and well-understood theoretical background for using these approaches, and their application in the existing limitations in real-world data such as a limited number of observations, or evaluation of partial/incomplete regimens.

With this, an important question remains to be addressed - among the currently available methods, which technique would be most appropriate in determining the effects of treatment from an incomplete exposure in the reduction of a health-related outcome using observational data?

In the Philippines, preterm birth remains to be a major health problem accounting for around two-thirds of all newborn deaths from respiratory-related conditions such as respiratory distress syndrome and neonatal pneumonia.[10] The pharmacologic agent corticosteroids, accelerate the maturation of the fetal lung and promote the production of pulmonary surfactant reducing the incidence of respiratory distress syndrome.[11]

Despite the recognition of antenatal corticosteroid (ACS) therapy, its utilization has not been optimistic attributed to the failure of clinicians to identify the onset of preterm labor,[12] and differing attitudes towards its use when completion of doses is not possible.[13] The current local guidelines recommend administering four doses of ACS to the pregnant woman intramuscularly at a twelve-hour interval.[14] Thus, some premature newborns were only able to receive only a single dose or not any dose of corticosteroids at all before delivery – resulting in poorer outcomes and prognoses among these neonates.

It has been noted that only 5% of women indicated to receive ACS, which account for 90% of childhood deaths globally in the year 2000. had the intervention among 42 countries.[15] An audit of Southeast Asian hospitals has recorded poor utilization of ACS in obstetric care,[16] the median rate of ACS utilization was 54%, the average rate in the Philippines was 47%, and 58% in Cambodia.[17]

Foreign guidelines mentioned the importance of administering the said single-dose regimen even when the second (or succeeding) doses cannot be given due to highly likely imminent, preterm delivery. However, there is still a lack of explicit recommendations in Philippine guidelines.[13]

The current study intends to use propensity score and machine-learning procedures to evaluate the utility of a single, 6-mg dose of Dexamethasone intramuscularly before birth which is given as an "emergency" dose among at-risk women. The focus of the current research is to demonstrate the estimation of treatment effects using various techniques.

METHODS

Before the conduct of the study ~~analysis,~~ approval from the University of the Philippines Manila – Review Ethics Board was ensured.

Participants and Setting

The dataset used was derived from a retrospective cohort of neonates delivered at 24 to less than 34 weeks of gestation in a tertiary general hospital in the Philippines.[18] The limited number of eligible patients in the institution where the dataset was obtained, and the transition of their medical records section have limited the availability of patient records (maternal and neonatal). However, the propensity score and machine-learning methods performed in the current study were developed to account for a limited number of observations, or few events per confounder.[1,19]

Identifiers and variables irrelevant to the current study were removed from the original data after permission was sought from the researchers who reviewed the patient charts.

Variables in the Dataset

The dataset included information from 562 patients with a disparate distribution of treated, those whoreceived a single-dose ACS (n = 439), and untreated, no ACS dose prior to delivery (n = 123), neonates in the study.

The exposure of interest is a dichotomous variable represented as the administration of a single-dose ACS or no dose given in utero. Likewise, the outcome of interest is also dichotomous, referring to the presence of any form of the respiratory-related condition among prematurely delivered neonates. These conditions included the presence of respiratory distress syndrome or transient tachypnea of the newborn, the need for surfactant administration; the need for oxygen support such as continuous positive airway pressure, oxygen hood, or mechanical ventilation; and the occurrence of respiratory failure and/or mortality from respiratory failure.

The association between the exposure and the outcome of respiratory-associated morbidity was determined using a large set of covariates. For the statistical approaches to control for bias and confounding, variables that are associated with the exposure alone, and those related to both the outcome and the exposure were included in the study dataset. These variables included maternal (e.g., maternal weight, poor obstetric history, presence of co-morbid conditions, maternal age), and neonatal covariates (e.g., birth weight, age of gestation during delivery, sex of the baby, Apgar scores) were accounted for in the analysis.

Statistical Analysis

The data analysis involved multiple steps, utilizing the software, Stata version 13.[20] Given the observation nature of the dataset, the clinico-demographic characteristics between the no ACS and the single-dose group are expected. Traditional regression models specifically crude, full multivariable, and adjusted logistic regression models were developed to estimate the association between exposure to ACS and respiratory morbidity.

Propensity scores (PS) were later conditioned and implemented and were estimated using conventional logistic regression and generalized boosting regression models. A causal diagram, as shown in Appendix C, was used to determine which variables were included in the propensity score (PS) estimation model. Other variables such as the birth weight, birth percentile, fifth minute Apgar score, and sex of the neonate were not included in the PS estimation model because they are more associated with the outcome, than the treatment assignment but considered in the measurement of effects.

The best-fitting logistic model to estimate a propensity score involved regressing the exposure (single-dose antenatal corticosteroids) with most of the known covariates. Initially, logit models that include interactions and higher order terms, to account for the complex pathways between the measured covariates, were planned to estimate the score. However, they were not able to achieve convergence, and the limited sample size led to a small matrix size unable to accommodate splines and interactions.

A generalized boosted regression algorithm was the other method utilized in estimating propensity scores. Most specifications used a maximum number of iterations (trees) at 10,000, a training fraction at 90%, a shrinkage factor of 0.001, using all observations for building each tree (bag: 1.0), and a depth of interaction set to five.

After estimating the PS, the achievement of balance on the covariates between the exposure groups was repeatedly assessed, and this iterative step is needed to ensure significant control of imbalance between the study groups. Plots such as kernel density, standardized bias, and paired bar charts were also used to examine visually whether assumptions in performing PS analysis were satisfied.

Once sufficient balance and overlap were achieved, these scores were implemented to ascertain the association between outcome and exposure. The estimated scores were to be implemented using matching models, and stratification for the logistic estimated scores. The boosted scores are implemented using inverse probability weighting and the approach proposed by Linden[21] where the scores were used both as weight and as a stratifying variable via marginal mean weighting through stratification. The difference in conditioning techniques is attributed to machine learning algorithms readily producing weights than the actual score, whereas an actual value can be derived from conventional PS estimation approaches.

Model fit and discrimination between the propensity score estimation models were assessed using the c-statistic, reduction in standardized differences, and the Brier scores. The best fitting model was selected using the following metrics: (a) width of the confidence interval of the odds ratios, (b) standard error of the regression, (c) Akaike information criteria AIC, and (d) Bayesian information criteria BIC.

These four metrics have been used in literature as important bases for model selection. The width of a confidence interval represents the amount of information collected and the precision with which the data are collected. Hence, a smaller width of the confidence interval is desired. While

Table 1. Comparison of model fit, prediction, and balance between estimated scores

Model	HL X2	p-value	Strata	MH X2	p-value	AUC (95% CI)	SE	Brier	Sanders	Reliability	Max ASD	ASD >0.10	Max VR
Logistic	8.11	0.44	9	2.78	0.95	0.71 (0.66-0.77)	0.028	0.1465	0.1499	0.0023	0.17	1/23	1.18
Boosting	9999		5	0.35	0.99	0.66 (0.61-0.71)	0.027	0.1674	0.1689	0.0067	0.20	8/23	1.41

HL – Hosmer-Lemeshow test, MH – Mantel-Haenszel test, AUC – Area under the receiver operating characteristic curve, SE – Standard error, Brier – Brier score, Sanders – Sanders modified Brier score, Reliability – Reliability-in-the-large, Max ASD – Maximum absolute standardized difference, ASD – Absolute standardized difference, Max VR – Maximum variance ratio

smaller values of the standard error are preferred suggesting a better fit between the values of the estimate and the true value of the unknown parameter.[22] The AIC aims at looking for the best-fitting approximate model, while the BIC focuses on identifying the most appropriate and parsimonious model.[23] A lower value of these information criteria suggests a closer approximation of the estimated model to the true data.

The best performing traditional, logit PS, and GBM PS models were used to estimate the average treatment effects on the sample (ATE), and the average treatment effects among neonates who were exposed to a single-dose ACS (ATET).

Moreover, Mantel-Haenszel bound estimates[24] were used on the estimates of effect to examine the possible impact of unobserved heterogeneity (hidden or residual bias) in the exposure groups.

RESULTS

Data from 123 (21.89%) maternal-neonatal dyads who did not receive any dose of ACS; and 439 (78.11%) who received a single dose of ACS before delivery were used to estimate the effects of the latter exposure. It has been established in a previous study[18] that there were inherent differences between the exposure groups (Appendix A) attributed to the lack of randomization and methodological constraints from observational study designs.

Moreover, the presence of indication bias can be suspected among women who are classified as high-risk pregnancies – like those with hypertension, congenital heart disease, or poor obstetric history are more commonly advised to receive the exposure (single-dose corticosteroids) as a rescue dose or an initial, incomplete dose – than other pregnant women do. Similarly, newborns from mothers who have preterm, premature rupture of membranes, or were born with other conditions – were also more likely to receive the exposure but at a higher risk of developing the outcome.[25]

As previously mentioned, the use of propensity score approaches is attractive considering the current dataset – observational, with too many potential confounders yet a relatively small number of observations. Propensity score analysis tends to control for confounding and biases by modeling the selection mechanism or allocation of treatment while at the same time, reducing the dimensionality of data by using a single scalar vector to estimate the association between the treatment and outcome of interest.[4]

Estimation of the Propensity Scores

Table 1 showed the goodness-of-fit statistic for the logistic estimated scores, and it demonstrated a satisfactory fit. Because of the lack of such tests in machine learning approaches, the optimal number of iterations was presented.

The optimal number of strata to examine associations was nine for the logistic estimation model while the optimal number of strata for the machine learning estimation model was five. The previous table also presented the area under the receiver operating characteristic (ROC) curve, and the computed overall Brier scores. The logit-estimated scores have higher values of the area under the curve, and lower values of the Brier scores – than scores estimated using generalized boosted models.

In terms of covariate balance, all covariates in either estimated score were within the (positive and negative) 20% estimate of bias after implementing them on the dataset. However, using a more stringent cut-off,[26] there were more covariates with ASD greater than 0.10 in the GBM-estimates scores than the logit ones. Other authors have suggested extending the cut-off up to 0.20, especially in smaller samples where not all covariates will likely have an estimated standardized difference within the previous threshold.[2]

Kernel density plots (Figure 1) were created to assess the degree of overlap, and how similar the distributions of the comparison groups are while accounting for the propensity score estimated. It showed a substantial region of common support indicating that the use of the estimated scores can emulate the counterfactual scenario.[27] However, the GBM-estimated scores demonstrated better overlap.

The paired bar charts (Figure 2) showed similarity between the distribution of the groups, and the logit-estimated scores, even with increasing the quantiles of the score to 18 subclasses, the plot showed that the unexposed participants had counterparts in the exposed group at all levels of the quantiles. While for the GBM-estimated scores, the plot showed that not all participants who received a single-dose ACS did not have unexposed counterparts at the fourth and fifth quantile. The presence of areas of non-overlap between the treatment groups has important implications in subsequent analysis to be performed. A remedy to these areas of no common support is to reinforce subjects within the overlap during the implementation of these scores.[28]

These findings and visual assessment suggest that the scores estimated from the logistic regression model can

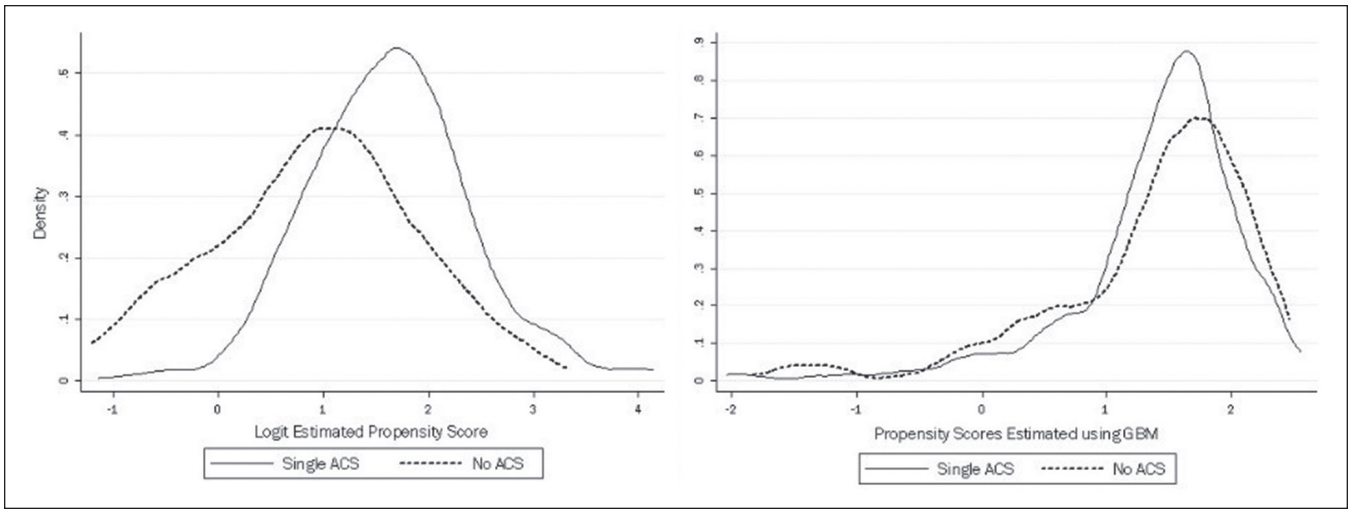


Figure 1. Kernel density plots of the estimated propensity scores.
ACS, antenatal corticosteroid

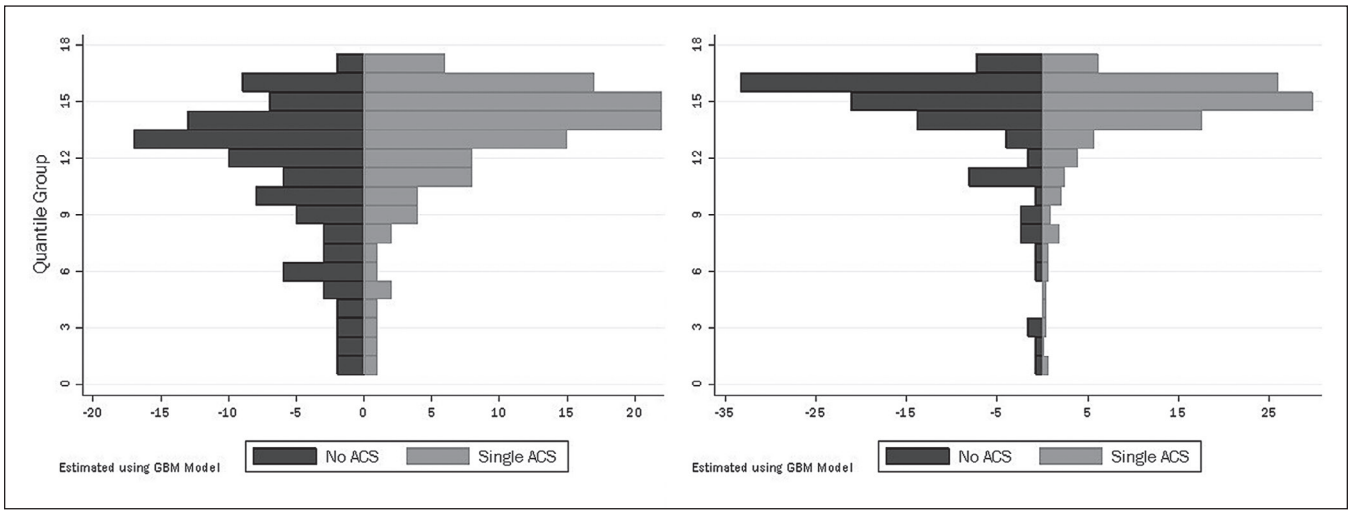


Figure 2. Paired bar charts of the estimated propensity scores.
ACS, antenatal corticosteroid

discriminate and forecast the antenatal corticosteroid status and achieve covariate balance between the comparison groups in the sample slightly better than those estimated using a machine-learning algorithm.

Building of Statistical Models

Several models were considered to examine the effects of exposure to a single-dose ACS on reducing respiratory complications among preterm neonates (Table 2). Traditional logistic regression models were developed to examine the association between ACS and respiratory-related outcomes. The crude logistic model (OR 0.43, 95% CI 0.27, 0.67, p < 0.01), and the full multivariable regression model (OR 0.45, 95% CI 0.23, 0.90, p = 0.02) suggested benefit from a single-dose ACS.

Likewise, the adjusted model formed after evaluating the presence of probable confounders and/or effect measure modifiers in the dataset demonstrated a similar reduction in the odds (OR 0.44, p = 0.01). The significant confounders accounted for in this model included the presence of a poor obstetric history, neonatal birth weight, and age of gestation during delivery. There were no effect measure modifiers or significant interactions identified in this model.

All propensity score-adjusted models utilized a "doubly robust" approach – where two implementation strategies were employed. The optimal (kernel) matching model was followed by covariance adjustment to address residual imbalances from the previous conditioning procedure. Only a few observations were left unmatched, and the proportion of the dyads not matched is similar between ACS groups. The

Table 2. Summary of the statistical models for single-dose antenatal corticosteroid (ACS) and respiratory morbidity

Models	N	OR	95% CI	SE	AIC	BIC
Crude logistic model	562	0.43	0.27-0.67	0.099	740.74	749.41
Full logistic model	562	0.45	0.23-0.90	0.159	439.22	525.85
Adjusted logistic model	562	0.44	0.23-0.84	0.145	434.52	456.18
PS matching model	546	0.37	0.17-0.80	0.147	326.72	352.54
PS stratification model	562	0.59	0.36-0.94	0.142	690.40	694.73
PS weighting model	562	0.51	0.31-0.85	0.131	1687.83	1696.50
PS MMWS model	562	0.44	0.28-0.70	0.103	741.50	750.27

OR – Odds ratio, SE – Standard error, AIC – Akaike information criterion, BIC – Bayesian information criterion, PS – Propensity score, MMWS – marginal mean weighting through stratification

matching model demonstrated a lesser likelihood of respiratory morbidity among those exposed to a single-dose ACS than the other statistical models (OR 0.37, $p = 0.01$).

The other model implementing the logit-estimated score was a stratification model (OR: 0.58, p: 0.03) where the stratum variable of the logit-estimated score into panel data, with conditional logistic regression, was performed after.

In terms of the PS estimated using boosting, an inverse probability to treatment (exposure) weighting followed by forcing the computed weights in a regression model was then implemented (OR: 0.51, $p < 0.01$). The weights were readily generated from the boosting algorithm, and the presence of big values was examined using an arbitrary cut-off at a value of twenty[29], but no "extreme" or "outlier" weights were identified. The said weights were then used also as a stratifying variable to create a marginal mean weighting through the sub-classification model (OR: 0.44, $p < 0.01$).

The effect measure from the IPTW model is more similar to those found when logit-estimated scores are used for stratification, and the one from the marginal mean weighting through stratification (MMWS) model is more alike to the odds ratio from the adjusted logistic model.

It was also noted that all the statistical models had point and interval estimates of the odds ratios less than the null value of one. Table 2 further showed that the different strategies concur the reduced likelihood of having respiratory morbidity among pre-term neonates can be attributed to exposure to a single-dose ACS given before delivery, than not receiving any dose at all.

The crude logistic had the smallest values of the confidence interval and the standard error of the model. However, a crude model is not advised for estimating treatment effects since it does not control for the presence of confounding variables. The PS-matching model had the best fit, parsimony, and adequacy compared to the other statistical models evidenced by having the smallest values for both Akaike and Bayesian information criteria.

Considering the observations in the previous section suggesting that that the estimated scores from logistic regression performed better than PS computed using a boosting approach, the PS matching model was deemed as the best fitting to estimate treatment effects conferred by a single-dose ACS over no ACS dose before preterm delivery.

Estimation of the Average Treatment Effects

Even if the PS matching model has been selected as the most appropriate model for estimating treatment effects, the researchers also considered estimated effects from the best performing conventional (adjusted logistic model) and boosting PS adjusted (MMWS) models (Table 3).

The negative values of either average treatment effects suggest improved outcomes among those who had a single-dose ACS, compared to the no ACS group. It also suggested a similarity to absolute risk reduction and can be used to compute other metrics such as numbers needed to treat.[30]

The effects of the boosting approach did not demonstrate a significant reduction in the likelihood of the outcome among those exposed to a single-dose ACS. Whereas the adjusted logistic and PS matching models indicated a relatively small but statistically significant reduction in the probability of the outcome in the sample, and among those treated.

Using the best-fitting model – the one utilizing PS matching, the average treatment effect alluded to a reduction in the risk of developing respiratory morbidity by around nine percent among preterm neonates who received a single-dose ACS, and a relatively smaller effect at eight percent among those who received the intervention.

Table 3. Assessment of the average treatment effects

Measure/Model	Effect (95% CI)	SE	z	p-value
ATE				
Adjusted	-0.27 (-0.18 to -0.37)	0.047	-5.78	<0.01
Matching	-0.09 (-0.03 to -0.15)	0.032	-2.84	<0.01
MMWS	-0.41 (-0.12 to 0.04)	0.039	-1.04	0.30
ATET				
Adjusted	-0.17 (-0.10 to -0.24)	0.037	-4.54	<0.01
Matching	-0.08 (-0.02 to -0.15)	0.034	-2.48	0.01
MMWS	-0.07 (-0.15 to 0.01)	0.042	-1.61	0.11

ATE – Average treatment effect, ATET – effect among those treated, MMWS – marginal mean weighting through stratification

Sensitivity Analysis

As previously mentioned, the use of different multivariable regression models and conditioning approaches on an estimated propensity score is, in itself, a form of sensitivity analysis.[31] Another approach is to compare the computed odds ratio against various likely and plausible measures of effect. Based on these comparisons, the selected propensity score-matching model tended to have an underestimated odds ratio from a single-dose ACS if we consider that the "true" effect lowers the likelihood of respiratory disease by half. However, the computed effect was not that much different from what was observed in other models.

The impact of unobserved confounding on the matching model was also evaluated using Mantel-Haenszel bound estimates (Appendix B). Assuming that there is no unmeasured confounding, the findings provide strong evidence that the use of single-dose ACS is associated with better neonatal outcomes (Γ: 1.00, QMH^{+}: 1.89, p = 0.03). However, if there is a hypothetical unmeasured covariate affecting the likelihood of the neonates to receive ACS beyond 20%, the odds ratio of the PS matching model will likely move toward the null value (Γ: 1.20, QMH^{-}: 1.14, p = 0.13).

It suggests caution in considering the estimated treatment effect from the PS matching as the description of a "definite true" effect from a single dose ACS on neonatal respiratory morbidity. However, it has been emphasized that the bound estimates are "worst-case" scenarios, and the contribution of the un-measured confounding variable on the exposure assignment would need to be quite large to undermine the estimate from the model.[32]

DISCUSSION

The general outcome from the models is that the use of single-dose antenatal steroids before delivery, though incomplete, showed a possible reduction in the risk of developing respiratory-associated conditions, compared to not receiving any steroid dose at all. These findings are consistent with studies,[33,34] showing that better outcomes can still be benefited from incomplete corticosteroid doses among preterm neonates. However, the estimated benefit from a single-dose ACS is relatively small compared to the more improved outcomes among those neonates who have completed the regimen.[12,16,17]

Observational studies play an important role when experimental studies are not yet available, or not possible to perform. Improvement in the design of the study or a focus on the statistical methodology has been recommended to generate better quantification of treatment effects derived from non-experimental studies.[1]

The appropriateness of using propensity score analysis was exemplified in the study since it can accommodate studies with binary or time-to-event outcomes, or when there are more confounders than can be adjusted realistically using conventional approaches.[35] Most studies have reported that propensity score analysis is a large sample strategy to satisfy the assumptions of balance and sufficient overlap[36] but there are no standard guidelines as to what constitutes enough sample size. However, the paradox of using propensity scores is that when using larger sample sizes, its statistical power was found to be lower than standard regression methods.[20]

The logistic PS estimation model was built using only first-order terms due to higher terms like splines and interactions were not accommodated. This contrasted with what has been advised in the literature.[37,38] Some authors mentioned that these terms result in a very complex model, or one heavily dependent on the functional forms of splines and interactions to control for bias and confounding.[39] The use of such terms in the PS model also does not assure that covariate balance is achieved.[40] However, this was not a concern for PS estimation models using machine-learning techniques since complex relationships are embedded in their inner workings.

Another unforeseen observation is the better performance of propensity scores estimated using conventional logistic regression compared to those derived from machine-learning algorithms. The logit-estimated scores had larger values of the area under the ROC curve, and lower values of the Brier scores and its decomposition values than boosting-estimated scores suggesting better capability to predict observed treatment status. It also had a more acceptable region of common support, even if the kernel density plot had better overlap in the GBM-estimated scores, the paired bar chart showed a lack of unexposed counterparts in some quantiles of the said score. In addition, there was better attainment of covariate balance from the standardized differences between the exposure groups in the logit-PS.

These findings were in contrast with what was found by several authors[7,41] where scores from traditional estimation approaches such as logit or probit regression had fewer desirable properties than those computed from ensemble methods like generalized boosting models. A possible explanation is that generalized boosted regression belongs to automated procedures using an iterative estimation strategy to optimize balance on select covariates.[42] Hence, the distribution of boosting-estimated scores is more similar between the single-dose and the no ACS group to mimic the randomization and allocation in experimental studies. However, the resulting pairings from the GBM-estimated scores may not be substantially alike to balance the potential confounding factors evidenced by the balance metrics found in the study.[43] This is given that boosting approaches tend to focus on improving variance metrics over covariate balance – but the achievement of the latter is more seen as an indicator of unbiasedness in propensity score literature.[44]

Another explanation would be its dependence on a large sample size since the data is reduced every time a predictor or covariate for estimation is considered. The trees created from the GBM-estimation model might have reached a stage

where there is not sufficient data to improve on the scores, yet the generated trees are still too shallow to achieve balance.[45] It has been noted that some algorithms may automatically prefer certain types of variables such as those with a greater number of categories without regard for their importance in prediction.[46]

In summary, it can be said that current evidence suggests and supports the use of logistic regression in estimating the propensity scores – because when the models are specified appropriately via consideration of interaction and higher order terms in the model. The previously mentioned disadvantages of the logistic regression appear to be reduced and preferred over the "black box" nature of machine learning algorithms for propensity score analysis.[47]

Another important finding is the better performance of propensity score matching, compared to other proposed methods of dealing with observational data. The advantage of propensity score matching over traditional methods would be the use of only one scalar variable, the probability of exposure, to create matches. Another would be the propensity score replacing the collection of identified covariates – even if some variables have a weak to moderate association with the outcome, the degree of possible harm can be neglected.

PS matching has relatively milder assumptions and drawbacks.[48] The use of a kernel matching procedure, a form of optimal matching, was appropriate for this dataset since it can deal with situations where there is a relatively small number of control (unexposed) subjects, or when there are few observations in the sample.[9] Moreover, the matching model used for the current study was doubly robust with regression adjustment addressing possible imbalance retained after performing the first PS conditioning technique.[37]

One downside of matching is the need for large samples to create good matches, especially because propensity score approaches are commonly used when either outcome or exposure is small.[49] Reduction in sample size, due to incomplete matching, can affect the external validity and avoid as much as possible. In the study, all the models except the matching model retained 562 observations in the analysis, but the loss in observations was smaller compared to other studies that used matching.[50] However, it has also been emphasized the trade-off between losing some data and improving the efficiency of the propensity score-matching model to generate more unbiased treatment effect estimates.[51] Moreover, the larger value of the standard error or interval estimate width is attributed to the preference for matching procedures to minimize covariate imbalance than variance-related metrics.[52]

In addition, the findings of the study did not show improvement in bias reduction and confounder control attributed to the use of weighting procedures. An explanation is that the generation of weights in propensity score analysis is prone to residual systematic error from the estimation models, thus reducing the advantages offered by sub-classification.[53] The use of weighting strategies to implement propensity scores was also discouraged in settings where the number of observations is small, just like the situation in the current study.[54]

The effects on those who received the exposure are slightly different from the effect in the overall sample. One possible explanation is that neonates exposed to a single-dose ACS in utero have larger birth weights and older ages of gestation during delivery. Hence, the better outcomes cannot solely be attributed to the exposure, but to prognostic variables – thus, the computed effect is smaller.

The current study showed that observational studies have a potential for use in quantifying the impact of certain interventions (e.g., treatments, programs, policies). This is important in situations like the current research inquiry where the conduct of an experimental or RCT design is not only unethical but also not feasible due to numerous constraints. The average treatment effects from the propensity score-matching model showed the degree of potential benefit from a partial dose of antenatal corticosteroid – considering the low proportion of individuals who can complete the ACS regimen before delivery.

A strength of the current study is the separation between estimating PS from two approaches, and then implementing these estimated scores using various strategies.[55] The presented measures of effect in the study were also more like to how the effects from experimental designs are presented such as average treatment effects, risk reduction, and numbers needed to treat.[56] These measures are also relatively more tangible and can lead to better planned clinical treatment and policy-based decisions.

In addition, the current study presented two ways of performing sensitivity analysis on the estimated treatment effects from propensity score-adjusted models. The first one simulated how different the estimated effect odds ratios varied given the range of plausible "true" values of the odds ratio for the association between single-dose antenatal steroids and respiratory morbidity.[57] Rosenbaum-based Mantel-Haenszel bound estimates were used as the second approach, to infer the possibility that unmeasured confounding variable/s introduce heterogeneity in the probability of exposure and/or the subsequent outcome.[58]

The inherent methodological constraints in the used dataset are an important limitation in the study, which included the relatively small number of observations, unequal distribution of treated and untreated observations, large numbers of covariates that are potential confounders, and the possible misspecification in the models. Propensity score analysis may control for some of these concerns but it does not assure a good remedy for systematic errors that are inherent when using non-experimental data.[59] Despite the increased use of PSA, another important limitation in the study was the lack of clear guidelines or best practices available in the literature that sets as a metric of determining if the PS procedure was performed well yet,[50] and more so, with machine learning procedures in the context of propensity score analysis.

In line with the said limitation, the analysis performed in the study was built on what was proposed in statistical references and reported in the available literature. It is recommended that future studies report sufficiently the various aspects and steps undertaken in performing the propensity score analysis – not just as a means of assessing the quality and validity of the findings, but also as a learning point for future investigators.

Future studies might also address specific concerns regarding the conduct of the propensity score analysis. One example is the use of calibration procedures to refine the estimation of propensity scores from covariate-heavy datasets while remedying problems such as large variance metrics, poor overlap, or incomplete control of bias.[45] Another is the evaluation of other balance-related metrics for machine-learning-based PS estimation,[41] and other validation procedures in the context of PS analysis.

Experimental designs are the ideal method of evaluating and estimating the effects of exposures such as treatments, programs, or policies. However, these observational studies play an important role in the creation of practice guidelines and policy development especially since experimental designs are too difficult or impossible to perform in the context of maternal and child health.[48]

The study findings suggest that improvement in the design of the study, and a focus on the statistical methodology have been recommended to generate better quantification of treatment effects derived from non-experimental studies.[5] Moreover, the study contributed to the possible array of statistical strategies to address bias, confounding, and other potential threats to the validity of effect estimates from these non-randomized data commonly collected in perinatal research.

However, propensity score-adjusted analysis and machine learning methods cannot replace the need for good study design, and quality of gathered data. These methods are also recommended as an adjunct approach to conventional regression analysis and not a replacement for such techniques.

Ultimately, the choice of which analytic technique highly depends on the objectives of the study, the size of the available data, the number of covariates and confounding variables identified as relevant; and the prevalence of the outcome and exposure.[60]

Acknowledgment

The authors acknowledge Dr. Mary Liezl Yu for allowing us to use the data gathered for her residency research paper entitled "Efficacy of single dose antenatal corticosteroid on reducing the morbidity and mortality of preterm infants: a retrospective cohort study" for this research activity.

Statement of Authorship

Both authors contributed in the conceptualization of work, acquisition and analysis of data, drafting and revising, and approved the final version submitted.

Author Disclosure

Both authors declared no conflicts of interest.

Funding Source

This study was funded by the authors.

REFERENCES

1. Shadish W, Steiner P. A Primer on Propensity Score Analysis. Newborn Infant Nurs Rev 2010; 10(1): 19-26.
2. Garrido M, Kelley A, Paris J, Roza K, Meier D, Morrison R, et al. Methods for Constructing and Assessing Propensity Scores. Health Serv Res. 2014; 49(5):1701-20.
3. Collins G, Le Manach Y. Comparing treatment effects between propensity scores and randomized controlled trials: improving conduct and reporting. Eur Heart J. 2012; 33(15):1867-9.
4. Guo S, Fraser M. Propensity Score Analysis: Statistical Methods and Applications. SAGE Publications, Thousand Oaks 2010: pp. 370.
5. Pan W, Bai H. Propensity score analysis: fundamentals and developments. The Guilford Press, New York 2015: pp. 402.
6. Cousens S, Hargreaves J, Bonell C, Armstrong B, Thomas J, Kirkwood B, et al. Alternatives to randomization in the evaluation of public health interventions: statistical analysis and causal inference. J Epidemiol Community Health. 2011; 65:576-81.
7. McCaffrey D, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods. 2004; 9(4): 403-425.
8. Luellen J. A Comparison of Propensity Score Estimation and Adjustment Methods on Simulated Data. ProQuest Information and Learning Company, 2007.
9. Luo Z, Gardiner J, Bradley C. Applying propensity score methods in medical research: pitfalls and prospects. Med Care Res Rev. 2010; 67(5):528-54.
10. Reolalas AT, Novilla MGM. Newborn deaths in the Philippines. Paper presented at 11th National Convention on Statistics (NCS), 2010; EDSA Shangri-La Hotel.
11. Briceño-Pérez C, Reyna-Villasmil E, Vigil-De-Gracia P. Antenatal corticosteroid therapy: Historical and scientific basis to improve preterm birth management. Eur J Obstet Gynecol Reprod Biol. 2019; 234:32-7.
12. Smith JM, Gupta S, Williams E, et al. Providing antenatal corticosteroids for preterm birth: a quality improvement initiative in Cambodia and the Philippines. Int J Qual Health Care. 2016; 28(6):682-8.
13. Philippine Obstetrical and Gynecological Society, Clinical Practice Guidelines on Preterm Labor and Preterm, Prelabor Rupture of Membranes, 2nd ed. Philippine Obstetrical and Gynecological Society (POGS) Foundation, Inc., Manila. 2010.
14. Committee on Obstetric Practice. American College of Obstetricians and Gynecologists Committee Opinion No. 713: Antenatal Corticosteroid Therapy for Fetal Maturation. Obstetrics and Gynecology. 2017; 130(2):e102-9.
15. Jones G, Steketee RW, Black RE, Bhutta ZA, Morris SS. How many child deaths can we prevent this year? The Lancet. 2003; 362(9377):65-71.
16. Pattanittum P, Ewens MR, Laopaiboon M, et al. Use of antenatal corticosteroids prior to preterm birth in four South East Asian countries within the SEA-ORCHID project. BMC Pregnancy Childbirth. 2008; 8(47):8.
17. Vogel JP, Souza JP, Gülmezoglu AM, et al. Use of antenatal corticosteroids and tocolytic drugs in preterm births in 29 countries: an analysis of the WHO Multicountry Survey on Maternal and Newborn Health. The Lancet. 2014; 384(9957):1869-77.
18. Yu M, Estrella A. Efficacy of a single-dose antenatal corticosteroid on reducing the morbidity and mortality of preterm infants: a retrospective cohort study. Philipp J Obstet Gynecol. 2015; 39(2):17-23.
19. Cepeda M, Boston R, Farrar J, Strom B. Comparison of logistic regression versus propensity score when the number of events is low

and there are multiple confounders. Am J Epidemiol. 2003; 158(3): 280-7.
20. StataCorp. Stata Statistical Software: Release 13. StataCorp LP, College Station 2013.
21. Linden A. Combining propensity-score based stratification and weighting to improve causal inference in the evaluation of health care interventions. J Eval Clin Pract. 2014; 20(6):1065-71.
22. Levy PS, Lemeshow S. Sampling of Populations: Methods and Applications. 4th ed: John Wiley & Sons, Inc.; 2008.
23. Acquah HDG. Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. J Dev Agric Econ. 2010; 2(1):001-006.
24. Becker S, Caliendo M. Sensitivity analysis for average treatment effects. Stata Journal 2007; 7(1): 71-83.
25. Hoxha A, Cnota W, Czuba B, Ruci A, Jarno M, Jagielska A, et al. A retrospective study on the risk of respiratory distress syndrome in singleton pregnancies with preterm premature rupture of membranes between 24+0 and 36+6 weeks, using regression analysis for various factors. Biomed Res Int. 2018: 7162478.
26. Haukoos JS, Lewis RJ. The propensity score. JAMA. 2015; 314:1637-8.
27. Bergstra S, Sepriano A, Ramiro S, Landewe R. Three handy tips and a practical guide to improve your propensity score models. RMD Open 2019; 5(e000953):1-4.
28. Fu E, Groenwold R, Zoccali C, Jager K, van Diepen M, Dekker F. Merits and caveats of propensity scores to adjust for confounding. Nephrol Dial Transplant. 2019; 34:1629-35.
29. Vittinghoff E, Glidden D, Shiboski M. Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, 2nd ed. Springer Publishing, New York 2012.
30. Austin P, Laupacis A. A tutorial on methods to estimating clinically and policy-meaningful measures of treatment effects in prospective observational studies: a review. Int J Biostat. 2011; 7(11):1-33.
31. Golinelli D, Ridgeway G, Rhoades H, Tucker J, Wenzel S. Bias and variance trade-offs when combining propensity score weighting and regression: with an application to HIV status and homeless men. Health Serv Outcomes Res Methodol 2012; 12(2-3):104-18.
32. Caliendo M, Hujer R, Thomsen SL. The employment effects of job creation schemes in Germany. Bonn, Germany: Institute for the Study of Labor; 2005. IZA Discussion Paper No. 1512.
33. Costa S, Zecca E, De Luca D, De Carolis M, Romagnoli C. Efficacy of a single dose of antenatal corticosteroids on morbidity and mortality of preterm infants. Eur J Obstet Gynecol Reprod. 2007; 131:154-7.
34. Elimian A, Figueroa R, Spitzer A, Ogburn P, Wiencek V, Quirk J. Antenatal corticosteroids: are incomplete courses beneficial? Obstet Gynecol. 2003; 102(2):352-5.
35. Guertin JR, Rahme E, Dormuth CR, LeLorier J. Head-to-head comparison of the propensity score and the high-dimensional propensity score matching methods. BMC Med Res Methodol. 2016; 16(22):1-10.
36. Dehejia R, Wahba S. Propensity score matching methods for non-experimental causal studies. Rev Econ Stat. 2002; 84:151-61.
37. McMurry T, Hu Y, Blackstone E, Kozower B. Propensity scores: Methods, considerations, and applications in the Journal of Thoracic and Cardiovascular Surgery. J Thorac Cardiovasc Surg. 2015; 150: 14-19.
38. Austin P. A tutorial and case study in propensity score analysis: an application to estimating the effect of in-hospital smoking cessation counseling on mortality. Multivariate Behav Res 2011; 46:119-51.
39. Song H. Assess Improvement of Balancing Covariates by Propensity Score approach using Generalized Boosted Model (GBM) and Application Based on National Cancer Database. Emory University, 2019. https://etd.library.emory.edu/concern/etds/k3569441r?locale=de
40. Krug G. Augmenting propensity score equations to avoid misspecification bias – Evidence from a Monte Carlo simulation. AStA Wirtsch Sozialstat Arch. 2017; 11:205-31.
41. Lee B, Lessler J, Stuart E. Improving propensity score weighting using machine learning. Stat Med 2010; 29(3): 337-346.
42. Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc Series B Stat Methodol. 2014; 76(1):243-63.
43. Goller D, Lechner M, Moczali A, Wolff J. Does the estimation of the propensity score by machine learning improve matching estimation? The case of Germany's programmes for long term unemployed. Labour Econ. 2020; 65:101855.
44. Li M. Using the propensity score method to estimate causal effects: a review and practical guide. Organ Res Methods. 2013; 16:188-226.
45. Ferri-Garcia R, Rueda M. Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. PLoS ONE. 2020; 15(4):e0231500.
46. Couronné R, Probst P, Boulesteix A. Random Forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinform. 2018; 19(270):1-14.
47. Ali M, Alhambra D, Lopes L, Ramos D, Bispo N, Ichihara M, et al. Propensity Score Methods in Health Technology Assessment: Principles, Extended Applications, and Recent Advances. Front Pharmacol 2019; 100(973): 1-19.
48. Gardner KC. Statistical methods for assessing treatment effects for observational studies. University of Louisville, 2014. https://ir.library.louisville.edu/cgi/viewcontent.cgi?article=1478&context=etd
49. Deb S, Austin P, Tu J, Ko D, Mazer C, Kiss A, et al. A review of propensity-score methods and their use in cardiovascular research. Can J Cardiol. 2016: 32:259-65.
50. Guo S, Fraser M, Chen Q. Propensity score analysis: recent debate and discussion. J Soc Soc Work Res 2020; 11(3):463-82.
51. Ho D, Imai K, King G, Stuart E. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Polit Anal. 2007; 15(3):199-236.
52. Colson K, Rudolph K, Zimmerman S, Goin D, Stuart E, van der Laan M, et al. Optimizing matching and analysis combinations for estimating causal effects. Sci Rep. 2016; 16(6):1-11.
53. Hong G. Marginal mean weighting through stratification: adjustment for selection bias in multilevel data. J Educ Behav Stat. 2010; 35: 499-531.
54. Raad H, Cornelius V, Chan S, Williamson E, Cro S. An evaluation of inverse probability weighting using the propensity score for baseline covariate adjustment in smaller population randomised controlled trials with a continuous outcome. BMC Med Res Methodol. 2020; 20(70):1-12.
55. Harder V, Stuart E, Anthony J. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods. 2010; 15(3):234-49.
56. Martens E, Pestman W, de Boer A, Belitser S, Klungel O. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol. 2008; 37:1142-7.
57. Groenwold R, Nelson D, Nichol K, Hoes A, Hak E. Sensitivity analyses to estimate the potential impact of unmeasured confounding in causal research. Int J Epidemiol. 2010; 39:107-7.
58. Litwok D. RMHBOUNDS: Stata module to refine mhbounds to remove the cap on the number of strata and replace the large sample approximation for E and V with exact moments. Statistical Software Components. 2020;S458807.
59. Thoemmes F, Kim E. A systematic review of propensity score methods in the social sciences. Multivariate Behav Res. 2011; 46(1):90-118.
60. Ross M, Kreider A, Huang Y, Matone M, Rubin D, Localio A. Propensity score methods for analyzing observational data like randomized experiments: challenges and solutions for rare outcomes and exposures. Am J Epidemiol. 2015; 181(12):989-95.

APPENDICES

Appendix A. Distribution of Clinico-Demographic Characteristics between Study Participants

Characteristics	No ACS (n=123) n (%)		Single-dose (n=439) n (%)	
	Without	with Outcome	Without	with Outcome
***Frequency** (%)*	30 (24.39)	93 (75.61)	189 (43.05)	250 (56.95)
Maternal Factors				
***Age during pregnancy** (years)*				
19 to 34	19 (63.33%)	70 (75.27%)	127 (67.20%)	172 (68.80%)
≤18 or ≥35	11 (36.67%)	23 (24.73%)	62 (32.80%)	78 (31.20%)
Education				
At most high school	18 (60%)	60 (64.52%)	140 (74.07%)	196 (78.40%)
Reached college	12 (40%)	33 (35.48%)	49 (25.93%)	54 (21.60%)
Maternal weight [a]	56.57 ± 10.44	54.62 ± 9.50	54.69 ± 9.65	54.66 ± 10.23
Gravidity [b]	3 (1–9)	2 (1–10)	2 (1–10)	2 (1–9)
Parity [b]	1 (0–6)	1 (0–8)	1 (0–9)	1 (0–8)
Prenatal visits				
≤4 consults	20 (66.67%)	68 (73.12%)	138 (73.02%)	189 (75.60%)
>4 consults	10 (33.33%)	25 (26.88%)	51 (26.98%)	61 (24.40%)
Poor obstetric history				
Absent	23 (76.67%)	75 (80.65%)	148 (78.31%)	213 (85.20%)
Present	7 (23.33%)	18 (19.35%)	41 (21.69%)	37 (14.80%)
History of abortion	7 (23.33%)	14 (15.05%)	23 (12.17%)	31 (12.40%)
Preterm labor	1 (3.33%)	6 (6.45%)	22 (11.64%)	13 (5.20%)
Presence of conditions				
Hypertension	1 (3.33%)	14 (15.05%)	24 (12.70%)	45 (18%)
Gestational diabetes	1 (3.33%)	3 (3.23%)	13 (6.88%)	10 (4%)
Placental conditions	-	1 (1.08%)	4 (2.12%)	5 (2%)
Other conditions	-	5 (5.38%)	5 (2.65%)	20 (8%)
Type of labor				
Spontaneous labor	24 (80%)	68 (73.12%)	124 (65.61%)	140 (56%)
Obstetric-induced	3 (10%)	12 (12.90%)	19 (10.05%)	49 (19.60%)
Medical-induced	3 (10%)	13 (13.98%)	46 (24.34%)	61 (24.40%)
Mode of delivery				
Vaginal	21 (70%)	62 (66.67%)	122 (64.55%)	141 (56.40%)
Abdominal	9 (30%)	31 (33.33%)	67 (35.45%)	109 (43.60%)

a – mean ± SD, b – median (range)
ACS – antenatal corticosteroids, GA – Gestational age

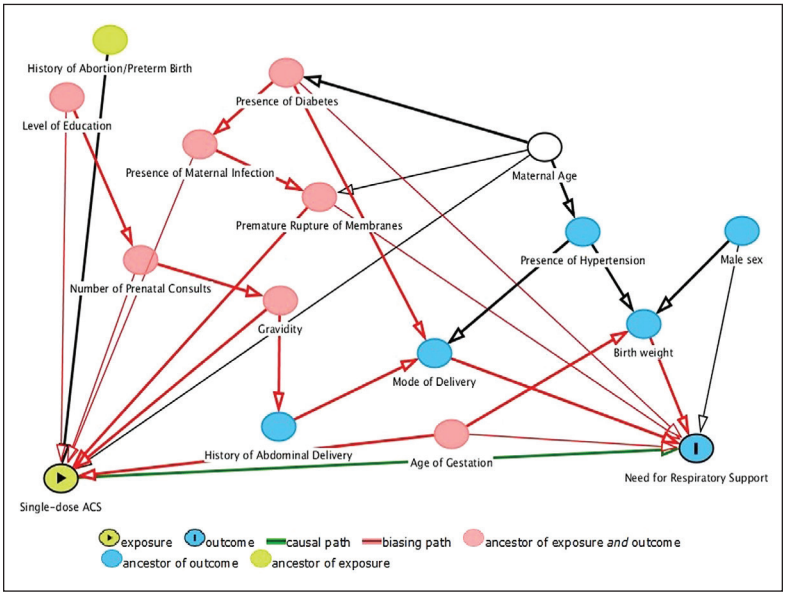
Appendix A. Distribution of Clinico-Demographic Characteristics between Study Participants *(continued)*

Characteristics	No ACS (n=123) n (%)		Single-dose (n=439) n (%)	
	Without	with Outcome	Without	with Outcome
Neonatal Factors				
Age of gestation *(weeks)*				
<28	-	25 (26.88%)	2 (1.06%)	32 (12.80%)
28 to 33	14 (46.67%)	51 (54.84%)	87 (46.03%)	151 (60.40%)
≤34	16 (53.33%)	17 (18.28%)	100 (52.91%)	67 (26.80%)
Birthweight *(grams)*				
<1500	15 (50%)	-	50 (26.46%)	5 (2%)
<2500	15 (50%)	38 (40.86%)	131 (69.31%)	136 (54.40%)
≥2500	-	55 (59.14%)	8 (4.23%)	109 (43.60%)
Rupture of membranes				
Yes	2 (6.67%)	6 (6.45%)	13 (6.88%)	32 (12.80%)
No	28 (93.33%)	87 (93.55%)	176 (93.12%)	218 (87.20%)
Sex of the neonate				
Female	11 (36.67%)	39 (41.94%)	89 (47.09%)	113 (45.20%)
Male	19 (63.33%)	54 (58.06%)	100 (52.91%)	137 (54.80%)
1st min Apgar score [b]	9 (5–9)	8 (2–9)	9 (6–9)	9 (1–9)
5th min Apgar score				
Poor (<7)	—	26 (27.96%)	—	22 (8%)
Good (7-9)	30 (100%)	67 (72.04%)	189 (100%)	228 (91.20%)
Term status				
Full term	21 (70%)	11 (11.83%)	77 (40.74%)	17 (6.80%)
Pre-term	9 (30%)	82 (88.17%)	112 (59.26%)	233 (93.20%)
Birth Percentile [b]	9 (2–91)	6 (1–97)	17 (2–97)	33 (23–40)
Pediatric aging category				
Small for GA	18 (60%)	56 (60.22%)	59 (31.22%)	127 (50.80%)
Appropriate for GA	11 (36.67%)	34 (36.56%)	124 (65.61%)	117 (46.80%)
Large for GA	1 (3.33%)	3 (3.23%)	6 (3.17%)	6 (2.40%)

a – mean ± SD, b – median (range)
ACS – antenatal corticosteroids, GA – Gestational age

Appendix B. Mantel Haenszel-based Rosenbaum Bound Estimates on the Matching Model

Gamma (Γ)	Q_{MH}^-	p-value	Q_{MH}^+	p-value
1.0	1.893	0.03	1.893	0.03
1.1	1.505	0.05	2.301	0.01
1.2	1.144	0.13	2.668	<0.01
1.3	0.813	0.21	3.009	<0.01
1.4	0.508	0.31	3.326	<0.01
1.5	0.224	0.41	3.624	<0.01
1.6	-0.041	0.52	3.905	<0.01
1.7	0.046	0.48	4.171	<0.01
1.8	0.279	0.39	4.423	<0.01
1.9	0.500	0.31	4.664	<0.01
2.0	0.710	0.24	4.894	<0.01



Appendix C. Causal Diagram for a Single-dose ACS and Respiratory-associated Morbidity.