Original Investigation

Using Propensity Score Modeling to Minimize the Influence of Confounding Risks Related to Prenatal Tobacco Exposure

Hua Fang, Ph.D.,[1] *Craig Johnson, M.A.,*[2,3] *Nicolas Chevalier, Ph.D.,*[2,3] *Christian Stopp, M.S.,*[2,3] *Sandra Wiebe, Ph.D.,*[2,3,4] *Lauren S. Wakschlag, Ph.D.,*[5] *& Kimberly Andrews Espy, Ph.D.*[2,3]

[1] *Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA*
[2] *Office of Research, University of Nebraska-Lincoln, Lincoln, NE*
[3] *Department of Psychology, University of Nebraska-Lincoln, Lincoln, NE*
[4] *Department of Psychology, University of Alberta, Alberta, Canada*
[5] *Department of Medical Social Sciences, Northwestern University, Chicago, IL*

Corresponding Author: Hua Fang, Ph.D., *Division of Biostatistics and Health Services Research, Department of Quantitative Health Sciences, University of Massachusetts Medical School, ACCES Bldg-7th Floor, Office: AC7-043, 55 Lake Avenue North, Worcester, MA 01655, USA. Telephone: 508-856-2502; Fax: 508-856-8010; E-mail: hua.fang@umassmed.edu*

Received January 25, 2010; accepted September 16, 2010

Abstract

Introduction: Despite efforts to control for confounding variables using stringent sampling plans, selection bias typically exists in observational studies, resulting in unbalanced comparison groups. Ignoring selection bias can result in unreliable or misleading estimates of the causal effect.

Methods: Generalized boosted models were used to estimate propensity scores from 42 confounding variables for a sample of 361 neonates. Using emergent neonatal attention and orientation skills as an example developmental outcome, we examined the impact of tobacco exposure with and without accounting for selection bias. Weight at birth, an outcome related to tobacco exposure, also was used to examine the functionality of the propensity score approach.

Results: Without inclusion of propensity scores, tobacco-exposed neonates did not differ from their nonexposed peers in attention skills over the first month or in weight at birth. When the propensity score was included as a covariate, exposed infants had marginally lower attention and a slower linear change rate at 4 weeks, with greater quadratic deceleration over the first month. Similarly, exposure-related differences in birth weight emerged when propensity scores were included as a covariate.

Conclusions: The propensity score method captured the selection bias intrinsic to this observational study of prenatal tobacco exposure. Selection bias obscured the deleterious impact of tobacco exposure on the development of neonatal attention. The illustrated analytic strategy offers an example to better characterize the impact of prenatal tobacco exposure on important developmental outcomes by directly modeling and statistically accounting for the selection bias from the sampling process.

Introduction

Unlike preclinical animal studies where confounding variables can be controlled and prenatal tobacco exposure can be assigned randomly as a treatment group, typical human exposure outcome studies use observational designs where sample likely differs in confounding variables between exposure groups. In statistical terms, unobserved selection bias exists, where exposure groups are not balanced as in a true experimental design. In fact, even with the best efforts of researchers to control for confounding variables using stringent sampling methods, unobserved selection bias typically exists (D'Agostino, 1998; Rosenbaum & Rubin, 1983, 1984). Ignoring selection bias can lead to unreliable or misleading estimates of causal effect that are the target of observational studies (Rosenbaum, 2002).

To address selection bias in observational studies and allow researchers to draw a causal inference from studies where randomization is not possible, an analytic method to control for selection bias is needed. Although there are several available, propensity score methods are being increasingly used. A propensity score is a probability value, estimated from confounding variables via a statistical model, for each subject who has the chance to belong to the "treatment" group (here those offspring who are tobacco exposed [TE]). In seminal work, Rosenbaum and Rubin (1983, 1984) showed that using propensity scores in hypotheses testing produced unbiased estimates of the true group difference. Unlike analysis of covariance, propensity score methods account for group differences by modeling the sampling process and addressing selection bias with a theoretically unlimited number of confounding variables related in any way to group selection (McCaffrey, Ridgeway, & Morral, 2004; Shadish, Cook, & Campbell, 2002; West, Biesanz, & Pitts, 2000).

doi: 10.1093/ntr/ntq170
Advance Access published on October 28, 2010

Once calculated, the propensity score can be included into statistical models as a single covariate, allowing researchers to statistically balance groups with less complex models and more statistical power (Braitman & Rosenbaum, 2002; Wang & Donnan, 2001).

To estimate propensity scores, most studies have used a parametric logistic regression model that assumes a specific underlying distribution and that the covariates are linear and additive on the log odds scale. Because covariates are usually non normal, nonlinear, and not additive, generalized boosted models (GBM; McCaffrey et al., 2004) that incorporate data mining and statistical techniques are a better alternative to calculate propensity scores (e.g., Friedman, 2001; Imbens, 2003). In data mining and machine learning literature, the term "boosting" refers to an algorithm that identifies the strongest model by building upon and learning from weaker models (Freund & Schapire, 1997; Friedman, 2002; Schapire & Singer, 1999). GBM expands boosting algorithms by using a collection of regression trees that outperform traditional approaches (Breiman, Friedman, Olshen, & Stone, 1984; Buhlmann & Yu, 2003; Friedman, 2002; McCaffrey et al., 2004). Compared with typical logistic regression, the appealing features of GBM include (a) using an automated data-adaptive modeling algorithm that can estimate the nonlinear relation between a variable of interest and a large number of covariates; (b) reduction in the chance of model misspecification and as nonparametric models, do not assume underlying distributions; (c) accommodation of various types of covariates (continuous, nominal, or ordinal) and missing values while allowing multicollinearity; (d) allowing estimated propensity scores to be used for covariate adjustment, weighting, matching, or stratification; (e) better balance of covariates, with fewer prediction errors; (f) and greater capability of removing bias in baseline differences between treatment and control groups.

The purpose of the present study is to demonstrate the application of this novel method, GBM, in a prenatal tobacco exposure study to test unobserved selection bias between TE and nonexposed (NE) neonates. In this study, exposure was measured prospectively, using self-report measures and bioassays during pregnancy. We selected emergent attention skills as the outcome from our earlier study (Espy et al., in press). Neonatal attention skills were measured three times during the first month of life. We hypothesized that propensity score modeling would account for substantial unobserved selection bias and that inclusion of the propensity score as a covariate would alter the pattern of prenatal tobacco exposure-related effects on early attention development. We also used birth weight, the most commonly reported outcome that is affected deleteriously by prenatal tobacco exposure (e.g., DiFranza, Aligne, & Weitzman, 2004), as a second exemplar outcome to test the efficacy of the propensity score method, where the inclusion of the propensity score as a covariate again was expected to reveal the magnitude of change of the exposure-related effects on birth weight.

Methods

Participants

Study flyers were distributed to pregnant women over a 4.5-year period at all obstetric and prenatal clinics at two sites in the Midwest: a rural five-county region and a small-sized city. Nine hundred and fifteen women contacted the laboratory and completed a screening interview to gather demographic information for selection and determine study eligibility (i.e., plan to deliver at a local hospital, speak English in the home, no binge drinking defined as ≥4 drinks per day, and no illegal drug use). Screened women who reported smoking around the last menstrual period (to capture smokers who underreport smoking during very early pregnancy and are often misclassified; England et al., 2007) or were actively smoking during pregnancy were recruited and enrolled. To reduce known demographic disparities between exposure groups, screened eligible nonsmokers were oversampled for enrollment based on Medicaid insurance status (a less intrusive proxy for income), race/ethnicity, and education (<14 years), resulting in 387 participants.

Participants completed a comprehensive adapted timeline followback interview during pregnancy at 16 weeks, 28 weeks, and just after delivery (termed 40 weeks hereafter). The interview gathered detailed information on smoking before and during pregnancy. Questions regarding use of alcohol and other substances, background, and health-related questions, such as diet, exercise, and medication use, were also included. A biological measure of tobacco exposure via cotinine levels was gathered for mothers and children using the DRI Cotinine Assay from U.S. Drug Laboratories. Mothers provided a urine sample at each interview during pregnancy, while neonatal cotinine was measured using a meconium sample taken from the neonate's diaper shortly after birth and urine samples at 2- and 4-weeks.

Despite our efforts to selectively focus on tobacco use and eliminate the confounding of illegal drug use through screening, 53 women denied use of marijuana during screening but admitted use on subsequent prenatal interviews or their neonate tested positive for marijuana at birth. We retained their data as it is not uncommon in prospective exposure studies for women to answer sensitive questions differently at screening than later during study enrollment when they are more comfortable. Because of the comorbidity of tobacco, alcohol, and marijuana use during pregnancy, particularly in heavier smokers, we elected to include a binary marijuana use variable in the propensity score estimation. However, due to the known large impacts on neonatal behavior that would mask prenatal tobacco effects, women/neonate data were excluded from eight participants with heavy drinking during any prenatal month (≥1 drink per day), 1 participant who was prescribed antipsychotic medication throughout pregnancy, and 17 participants who were born ≤35-week gestation.

Procedures

Prenatal Tobacco Exposure Measurement and Group Classification

Prenatal tobacco exposure was determined using the number of maternal self-reported cigarettes during prenatal smoking and biospecimen assays in a three-step process. First, women who self-reported smoking any cigarettes during the prenatal period on any maternal prenatal interview were classified initially as TE and those who reported no smoking during the period on all interviews were classified initially as NE. Then, the consistency of self-reported smoking behavior across interviews was examined for congruence with initial group assignment. Where smoking status was consistent across interviews and agreed with the last smoking date, the exposure group assignment remained. If these criteria were not met, the reported last smoking dates

across the interviews were examined relative to the last menstrual period. If a participant was initially classified NE despite last smoking dates falling in the window of pregnancy, that participant was reclassified as TE. Using this procedure, 16 participants were reclassified. Finally, the results of the biospecimen sampling were considered, as self-reported smoking can underestimate true maternal smoking due to social undesirability (Pley et al., 1991). Using the cutoff value recommended by U.S. Drug Laboratories, two women with urine cotinine values >100 ng/ml were reclassified as TE. Among the 361 neonates, 189 were classified as TE and 172 as NE.

The exposure group variable, determined by maternal self-reported cigarette smoking and cotinine levels, reflected the direct group-level effect of tobacco exposure incurred by the neonate during pregnancy. To capture the effects of second hand environmental smoke exposure to the mother that contributes indirectly to offspring exposure, the self-reported number of smokers in home during pregnancy and daily partner smoking amount in the presence of the participant (average value across the 16-, 28-, and 40-week interviews) were included as predictors in these models.

As expected, the mean cotinine levels in maternal urine and neonate meconium differed among the TE and NE groups at all timepoints (all ps < .01). The mean TE maternal urine cotinine was 364.95 ng/ml at 16 weeks and 333.21 ng/ml at 28 weeks. Mean NE maternal urine cotinine was 5.70 ng/ml at 16 weeks and 10.75 ng/ml at 28 weeks. At the 40-week interview, the mean maternal urinary cotinine level for the TE group dropped to 75.7 ng/ml, whereas for NE women remained unchanged (11.69 ng/ml). The mean cotinine level in infant meconium of the TE (196.19 ng/ml) was significantly higher than the NE group (0.63 ng/ml, p< .001).

Outcomes

Neonates were administered a standardized neonatal temperament assessment (NTA; Riese, 1982, 1986) three times in the neonatal period: approximately two days after birth in the hospital (called at birth hereafter), 2 weeks in a university laboratory, and 4 weeks in the participant's home. The NTA has demonstrated reliability (Riese, 1986), and 4% of all assessments were coscored and yielded mean interrater module reliabilities between .89 and .99. Individual NTA items were treated as multiple behavior indicators of three latent constructs that were identified empirically using principal-components analysis (Espy et al., in press); Attention/Orientation (AT), capturing infants' responses to auditory and visual stimuli and overall degree of alertness; Irritable Reactivity, summarizing infants' irritability during orientation items and reflex elicitation procedures; and Stressor Dysregulation, reflecting infants' latency to soothe after the cold disc and pacifier withdrawal stress tests. Espy et al. (in press) provide further details related to NTA administration and data reduction. Although three latent constructs captured neonatal behavior, substantive exposure effects were noted mainly for the AT factor score in the Espy et al. (in press). Thus for the purposes here, only the AT domain was examined. For the second outcome, weight at birth, the neonate's birth weight in grams as recorded by the hospital staff at delivery, was used as the dependent variable.

Analysis

Using propensity scores in analyses requires three basic steps: (a) propensity score estimation, (b) hypotheses testing with and without propensity score adjustment, and (c) sensitivity analysis. Each step is described in detail in the following sections.

Step 1: Propensity Score Estimation

In theory, an unlimited number of confounding variables can be considered and included in propensity score estimation. These variables do not have to be related to one another and can be continuous or categorical variables. However, all included confounding variables should have a theoretical rationale for inclusion.

Smoking during pregnancy cooccurs with numerous potential confounding variables that are related to childhood outcomes, including maternal psychiatric symptoms of hostility, depression (Anda et al., 1990; Fergusson, Goodwin, & Horwood, 2003; Rodriguez, Bohlin, & Lindmark, 2000; Schuetze & Eiden, 2006; Whiteman, Fowkes, Deary, & Lee, 1997), and anxiety (Parton et al., 1998), and Attention-Deficit Hyperactivity Disorder (ADHD) (Flick et al., 2006; Goodwin, Keyes, & Simuro, 2007; Kodl Middlecamp & Wakschlag, 2004). Pregnant smokers are also more likely to be young, poor, unmarried, and engage in other risky health behaviors during pregnancy, including alcohol and other drug use, and have suboptimal nutrition (Baghurst, Tong, Woodward, & McMichael, 1992; Breslau, 1995; Dani & Harris, 2005; Pickett, Wilkinson, & Wakschlag, 2009). Therefore, in this study, we gathered information pertaining to these maternal background variables through comprehensive interviews during pregnancy at 16-, 28-, and 40 week. Table 1 provides the maternal variables collected, which included demographic information, healthy diet (calculated by an average score of each subject across three visits if consumption of tuna, fish, bread, fruit, vegetables, and dairy were reported [yes/no]), mother's weight, prenatal alcohol use (drinks per day per month), prenatal marijuana use (yes/no), and prenatal prescription medication (yes/no for each medication). In addition to the interviews, during the 28-week session, participants completed the Brief Symptom Inventory (Derogatis, 1993) to assess maternal psychopathology symptoms and the Connors Adult ADHD Rating Scale—Short Form (Connors, Erhardt, & Sparrow, 1998) to measure ADHD symptoms. Mothers completed the Woodcock–Johnson Brief Intellectual Ability assessment during the 44-week postnatal interview to measure general intelligence (Woodcock, McGrew, & Mather, 2001). Standardized scores derived from instrument normative tables were used in all analyses. Less than 3% of the data were missing for the included confounding variables. Table 1 provides the 42 potential confounding variables means or proportions by exposure group.

A propensity score was calculated for each participant using the 42 confounding variables and the GBM-based "twang" package in R 2.8.1 (Friedman, 2002; McCaffrey et al., 2004; R Development Core Team, 2008; Ridgeway, 2006).

Step 2: Hypotheses Testing of Exposure Effect On AT and Birth Weight With and Without Propensity Score Adjustment

Hypothesis testing was conducted to determine if the exposure effect estimated from a statistical model would increase, decrease, or remain the same after controlling for selection bias. The obtained propensity score (single propensity score covariate), the exposure grouping variable (predictor of interest), and two maternal secondhand smoke exposure variables were entered into a latent multiple indicator growth model for neonatal attention skills (MIGM, performed in Mplus 6.0;

Muthén & Muthén, 2007; see Supplementary Figure 1) and a linear regression model for infant birth weight.

Neonatal attention scores and weight at birth were the respective outcome variables. For neonatal attention, the multiple indicator growth model that characterized developmental change in AT scores across age was used. This model integrates the structural equation approach of the relations between observed behavior indicators and latent constructs (e.g., NTA visual stimuli items to the AT construct) with the multilevel model conceptualization of age (at birth, 2, and 4 weeks) within subjects (Muthén & Muthén, 2007). Measurement invariance was specified and tested by holding the intercepts and factor loadings of the indicators equal across age. The maximum likelihood estimator with robust *SE*s (MLR) was used to allow for missing data at random as well as nonnormal and nonindependence outcomes (Yuan & Bentler, 2000). For the MLR estimator, the chi-square likelihood ratio test based on log likelihood values and scaling correction factors (Satorra, 2000) was used with Akaike's information criterion (AIC) and Bayesian information criteria (BIC) to examine model fit. The residual variances of the factor indicators (i.e., individual items) and the latent factors were estimated and allowed to differ across age. The regression models were used to estimate birth weight with TE/NE exposure grouping variable and two maternal secondhand smoke exposure variables as predictors.

Step 3: Sensitivity Analysis

Using propensity scores helps examine the influences of measured confounding variables on exposure effects, although no study can measure all the possible confounding influences. The inability to include all potential confounding variables can result in hidden bias for estimated effects (Rosenbaum, 2002). In this study, sensitivity analyses were performed in R 2.8.1 (Friedman, 2002; McCaffrey et al., 2004; R Development Core Team, 2008; Ridgeway, 2006). To begin, one observed confounding variable was removed from the propensity score model, treating it as an unobserved variable, and then the propensity score recalculated. Next, an obtained ratio of propensity scores with and without this confounding variable was computed for each person (McCaffrey et al., 2004; Ridgeway, 2006; Rosenbaum, 2002). This confounding variable then was added back into the model, the next confounding variable removed, and the process repeated. Finally, a worst-case scenario was repeatedly simulated for each removed confounding variable to reexamine the exposure effects on developmental parameters (i.e., intercepts, linear slopes, and quadratic decelerations for AT scores from our growth models) and birth weight. The worst-case scenario assumes a larger and more unlikely relation between the developmental parameters and calculated ratios than the actual observed correlations. In this study, an absolute correlation of .99 was used to illustrate this highly unlikely circumstance. If the worst-case scenario resulted in dramatically different model estimates, then exposure effects were considered susceptible to hidden bias (McCaffrey et al., 2004). That is, the estimate exposure effects may be dramatically affected by latent confounding variables.

Results

Propensity Score Estimation

The relative influence, or percentage increase in the logistic log likelihood (Friedman, 2001), of each confounding variable was obtained from the GBM. Relative influence, provided in the rightmost column of Table 1, indicates a variable's contribution to estimating the propensity score. The rank among confounding variables was created according to the degree of relative influence, with the higher the contribution, the more important the confounding variable is to propensity score calculation. Results showed maternal alcohol use during first month of pregnancy, education, and alcohol use around conception as being the three most influential variables. It is important to note that we cannot conclude or infer any relationship between any confounding variable and outcomes through the propensity scores. The propensity score approach in hypothesis testing is only used to balance compared groups, reduce the selection bias for a specific sample, and help reveal the more accurate exposure effect, regardless of the relation among confounding variables and outcomes. As shown in Table 1, 26% of the increase in model likelihood was due to alcohol-use variables, 28% to maternal mental health variables (e.g., maternal depression, anxiety, hostility, inattention, impulsivity, and hyperactivity), 26% to demographics (e.g., marital status, age, education, intelligence, ethnicity, insurance status, and number of pregnancies), and 20% to maternal health variables. Figure 1 displays the distribution of calculated propensity scores by exposure groups. The large difference between the TE and NE groups indicates that selection bias exists despite the stringent sampling plan used to reduce confounding influences. Based on these results, the propensity score variable was included as a covariate in the multiple indictor growth model for AT and in the regression model for weight at birth.

Hypotheses Testing With and Without Propensity Score Adjustment

Attention

Smaller AIC and BIC and significant MLR chi-square likelihood ratio tests indicated that the quadratic model (AIC = –4,517.59; BIC = –4,311.48; $\chi^2_{\_MLR\ difference} = 28.76$, $p < .01$) fit better than the linear model (AIC = –4,302.90; BIC = –4,116.23). These three indices (AIC = –2,101.00; BIC = –1,872.38; $\chi^2_{\_MLR\ difference} = 24.89$, $p < 0.05$) also indicated that the full model including the propensity score fits the data better than the model without propensity scores (AIC = –1,882.20; BIC = –1,665.21). The calculated developmental trajectories of the AT scores across age by exposure groups are plotted in Figure 2. Centering at 4 weeks of age, the growth models without a propensity score showed that the intercept and linear change rate of TE neonates did not differ from their NE peers ($\gamma_{\_intecept} = 0.016$, $SE = 0.018$, $p = 0.39$; $\gamma_{\_slope} = 0.010$, $SE = 0.017$, $p = 0.58$). The exposure groups also did not differ in their quadratic deceleration rate ($\gamma_{\_quadratic} = –0.001$, $SE = 0.004$, $p = 0.77$) over the first month of their life. The two maternal secondhand exposure measures were not related to neonatal attention growth (number of self-reported smokers in home during pregnancy, ps > 0.10, and daily partner smoking amount in the presence of the participant, ps > 0.30). With propensity scores included, growth models showed that neonatal attention differences between TE and NE were larger in magnitude. Furthermore, compared with NE peers, TE neonates score marginally lower in AT at 4 weeks of age ($\gamma_{\_intecept\_ps} = –0.042$, $SE = 0.027$, $p = 0.10$), with a marginally slower linear change rate ($\gamma_{\_slope\_ps} = –0.041$, $SE = 0.023$, $p = 0.08$) and marginally greater quadratic deceleration rate ($\gamma_{\_quadratic\_ps} = –0.009$, $SE = 0.006$, $p = 0.10$) over

Table 1. Descriptive Statistics By Exposure Group and Relative Influence in Propensity Scores

	Tobacco exposed		Nonexposed			
Confounding variables	*M*/%	*SD*	*M*/%	*SD*	Rank[a]	%
Maternal age at delivery (years)**	25.2	4.9	26.6	4.9	4	7.65
Maternal education (years)***	12.98	1.56	13.88	1.71	2	9.83
% Medicaid	85	—	84	—	31	0.26
% Married***	37	—	57	—	32	0.25
Maternal race (% White)	77	—	77	—	28	0.34
Maternal weight						
Prepregnancy	162.2	48.2	167.2	45.8	9	3.61
Delivery	197.7	47.9	196.6	45.4	16	2.05
Gain**	35.5	19.6	29.4	14.9	8	4.38
Number of previous pregnancies	1.68	2.05	1.61	1.55	20	1.44
Healthy diet	4.38	0.69	4.51	0.76	13	2.52
Exercise (% three times per week)						
Prepregnancy	47	—	53	—	26	0.54
16 weeks	39	—	48	—	21	1.33
28 weeks	42	—	42	—	35	0.11
Delivery	31	—	34	—	34	0.17
% Prenatal marijuana use***	20	—	5	—	7	4.45
Average number of alcohol drinks per day+						
At last menstrual period***	0.467	0.926	0.109	0.337	3	8.55
Month 1 pregnancy***	0.245	0.398	0.036	0.092	1	14.88
Month 2 pregnancy**	0.032	0.114	0.005	0.034	23	1.22
Month 3 pregnancy	0.006	0.037	0.002	0.008	33	0.23
Month 4 pregnancy	0.003	0.010	0.001	0.007	39	0.03
Month 5 pregnancy	0.003	0.012	0.001	0.006	29	0.32
Month 6 pregnancy*	0.004	0.016	0.001	0.006	30	0.30
Month 7 pregnancy*	0.005	0.018	0.001	0.005	40	0.03
Month 8 pregnancy	0.005	0.027	0.001	0.010	38	0.05
Month 9 pregnancy	0.005	0.028	0.001	0.007	41	0.02
% Prenatal prescription medication						
Antidepressants	12	—	8	—	36	0.10
Opioid-based analgesics	22	—	19	—	27	0.47
Asthma	5	—	7	—	37	0.06
Thyroid**	2	—	4	—	—	0.00
BIA maternal IQ score***	95.08	11.52	99.51	12.20	5	6.43
BSI subscale T score						
Anxiety*	50.70	9.61	48.59	9.21	19	1.54
Depression**	53.70	8.74	51.21	9.00	17	1.94
Hostility	57.49	9.16	56.33	8.54	18	1.88
Interpersonal sensitivity	53.28	9.12	52.95	9.16	15	2.08
Obsessive–compulsive	56.91	10.73	56.83	10.18	22	1.31
Paranoid ideation	52.59	9.13	51.24	8.81	25	0.88
Phobic anxiety**	51.62	8.52	49.35	7.24	11	3.10
Psychoticism**	55.40	9.49	52.67	8.73	24	0.98
Somatization	58.95	8.70	58.23	8.46	14	2.47
CAARS subscale T score						
Hyperactivity	47.98	8.32	46.46	7.33	6	5.81
Impulsivity	46.16	6.92	45.66	6.94	12	2.78
Inattention	48.03	8.40	47.83	7.99	10	3.60

Note. BIA = Woodcock-Johnson III Brief Intellectual Ability; BSI = Brief Symptom Inventory; CAARS = Connors Adult ADHD Rating Scale-Short Form.

[a]Rank based on the magnitude of the relative influence listed in the last column; rank is not given if the relative influence is 0%.

*$p < .05$; **$p < .01$; ***$p < .001$.

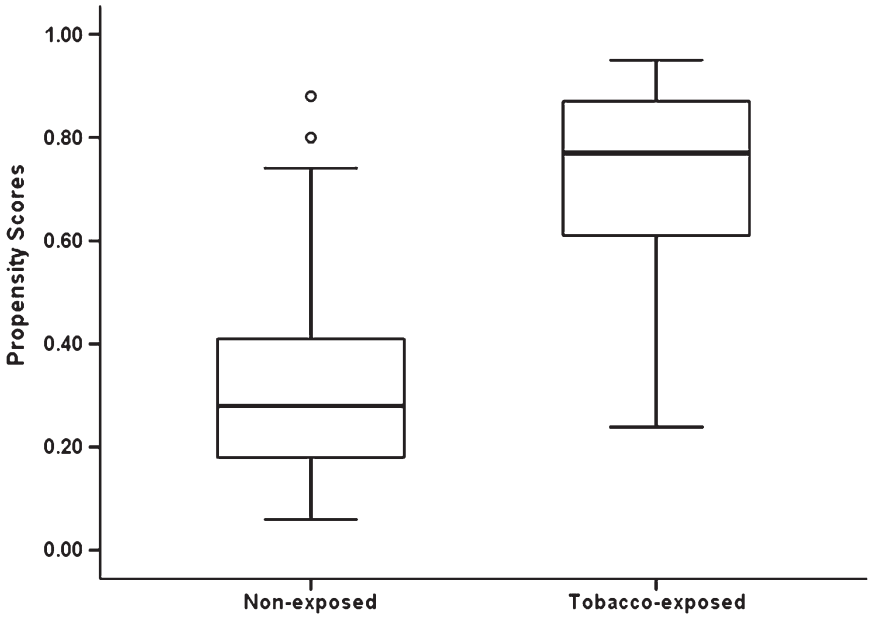

Figure 1. Propensity score distribution by tobacco exposure group status (the two dots indicated two nonexposed [NE] have relatively high propensity scores and the rest of NE neonates have propensity scores below .40).

the first month. Again, maternal secondhand smoke exposure variables were unrelated to attention growth (number of self-reported smokers in home during pregnancy, *p*s >0.12, and daily partner smoking amount in the presence of the participant, *p*s > 0.32).

Birth Weight

Similar to the AT results without propensity scores included, TE and NE groups did not differ in birth weight ($\gamma_{\_bwt} = -71.352$, $SE = 49.826$, $p = 0.15$). However, inclusion of the GBM estimated propensity scores, the weight difference between the two exposure groups, was greater in magnitude and reached marginal statistical significance ($\gamma_{\_bwt} = -133.309$, $SE = 73.371$, $p = 0.07$).

Sensitivity Analysis

The Supplementary Table selectively presents resulting prenatal exposure effects on the AT developmental parameters under the worst-case scenario after removing each of the top five influential confounding variables (as indicated in Table 1). These results indicated that the prenatal tobacco exposure effect did not appear to be sensitive to hidden bias as the worst-case scenario did not result in any dramatic change in the exposure effect on these developmental parameters. The same procedures were used to examine the hidden bias for the exposure effect on birth weight and again with no hidden bias found.

Discussion

The purpose of this study was to test the presence and evaluate the impact of selection bias in a carefully selected prospectively recruited observational sample. We then applied a GBM model to derive a propensity score for each individual. Using the derived propensity score as a covariate, hypothesis testing was

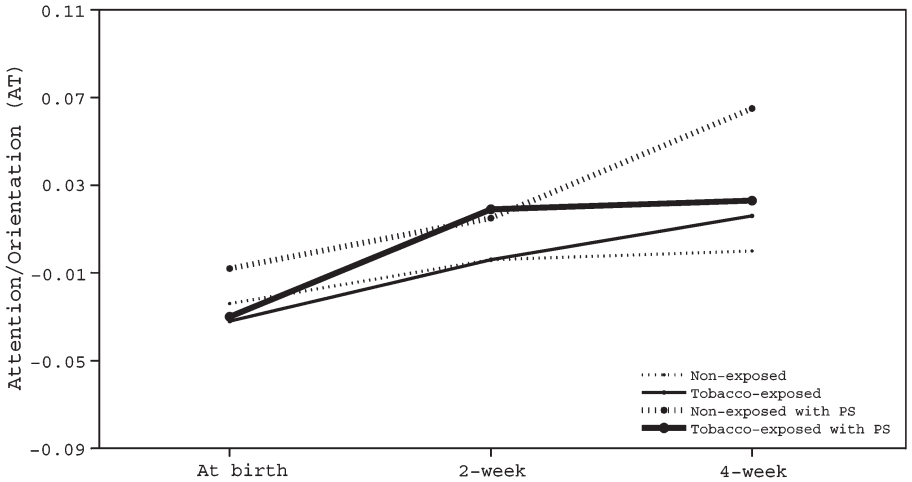

Figure 2. Exposure-related group differences in Attention/Orientation (AT) with and without propensity score adjustment.

conducted to determine if there were changes in the effects of prenatal tobacco exposure on important outcomes when propensity score covariate was included. Without propensity scores, the TE and NE groups did not differ in orientation to, and attentive tracking of, auditory and visual stimuli or in weight at birth. However, with a propensity score covariate included in the models, the exposure-related effects were larger in magnitude. In comparison with NE neonates, those exposed had lower attention and linear change rate at 4 weeks of age, a greater deceleration in attention skills over the first month of life, and weighed less at birth. These attention differences observed at 4 weeks of age, well after direct prenatal exposure has ceased, were not apparent in other studies when other analyses of covariance methods are used (Yolton et al., 2009). Similarly, the inclusion of propensity scores helped uncover the exposure group-level differences in birth weight that are not always evident in modern tobacco studies where the amount of smoking is substantially lower than studies conducted in earlier decades (Lumley, 1987; Shiono, Klebanoff, & Rhoads, 1986). Without the inclusion of propensity scores, the selection bias related to unaccounted background variables appears to have obscured exposure-related differences in neonatal attention and weight at birth. Of course, a different result might be obtained for other outcomes, for example, the Irritable Reactivity or Stressor Dysregulation domains from the NTA that were not examined here.

Although the statistical significance of tobacco exposure effect "improved" with the inclusion of the propensity score, that is not the purpose of propensity score modeling. Rather, propensity scores are included to minimize and theoretically eliminate selection bias related to confounding variables, thereby helping reveal the more accurate exposure effects. Comparing the results of the statistical models without and with the propensity scores, there are three possible results, that is, the magnitude of exposure effect can increase, decrease, or remain about the same. Larger or smaller exposure effects indicate that selection bias exists and needs to be tested to better characterize true exposure effects. Effects that are similar with and without propensity scores indicate that selection bias likely is negligible, which is also an important insight. Regardless of magnitude and direction of differences, this study indicates that section bias existed despite careful selection procedures used to minimize difference in background variables, as is common in modern observational designs for human teratological investigations. Propensity score modeling offers the opportunity to account for selection bias and thereby provide a more accurate and complete interpretation of statistical results. However, one disadvantage of propensity score approach is that the propensity scores are calculated by treating exposure group as a categorical variable (and cannot be computed directly on a continuous exposure variable), which might lead to some loss of information.

Taken as a whole, our findings illustrate three key points. First, the GBM method captured the selection bias and enhanced estimation of the influence of prenatal tobacco exposure on neonatal attention and birth weight. Second, despite careful and prospective selection methods, the influences of confounding variables appeared to dilute exposure-related differences in the development of early attention/orientation skills, as well as in birth weight, between TE and NE neonates. Third, because of the influence of selection bias, exposure-related outcome differences reported previously in other studies may be misattributed in magnitude and/or direction. Incorporating the propensity score methods illustrated here into the modeling strategy offers one potential method to better characterize the true impact of prenatal tobacco exposure on important developmental outcomes in observational studies by statistically accounting for selection bias related to confounding influences.

Supplementary Material

Supplementary Table 1 and Figure 1 can be found at *Nicotine and Tobacco Research* online (http://www.ntr.oxfordjournals.org/).

Funding

This research was supported in part by National Institutes of Health (R01 DA014661, 2003–2008; DA023653, 2009–2013; DA024769, 2008–2010; MH065668, 2004–2014, and HD050309, 2006–2011).

Declaration of Interests

The authors have no competing interests related and had access to all relevant data.

Acknowledgments

The authors acknowledge the participating families, hospital staff, and project personnel who made this work possible.

References

Anda, R., Williamson, D., Escobedo, L., Mast, E., Giovino, G., & Remington, P. (1990). Depression and the dynamics of smoking: A national perspective. *Journal of the American Medical Association*, *264*, 1541–1545.

Baghurst, P., Tong, S., Woodward, A., & McMichael, A. (1992). Effects of maternal smoking upon neuropsychological development in early childhood: Importance of taking account of social and environmental factors. *Paediatric and Perinatal Epidemiology*, *6*, 403–415.

Braitman, L., & Rosenbaum, P. (2002). Rare outcomes, common treatments: Analytic strategies using propensity scores. *Annals of Internal Medicine*, *137*, 693–695.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Breslau, N. (1995). Psychiatric comorbidity of smoking and nicotine dependence. *Behavior Genetics*, *25*, 95–101.

Buhlmann, P., & Yu, B. (2003). Boosting with the L2 loss: Regression and classification. *Journal of the American Statistical Association*, *98*, 324–339.

Connors, C. K., Erhardt, D., & Sparrow, E. (1998). *CAARS—Self-report: Short version (CAARS—S: S)*. North Tonawanda, NY: Multi-Health Systems.

D'Agostino, R. (1998). Propensity score methods for bias reduction in the comparison of treatment to a non-randomized control group. *Statistics in Medicine*, *17*, 2265–2281.

Dani, J., & Harris, R. (2005). Nicotine addiction and comorbidity with alcohol abuse and mental illness. *Nature Neuroscience*, *8*, 1465–1469.

Derogatis, L. R. (1993). *Brief Symptom Inventory (BSI): Administration, scoring and procedures manual*. Minneapolis, MN: NCS Pearson.

DiFranza, J. R., Aligne, C. A., & Weitzman, M. (2004). Prenatal and postnatal environmental tobacco smoke exposure and children's health. *Pediatrics*, *113*, 1007–1015.

England, L. J., Grauman, A., QainWilkins, D. G., Schisterman, E. F., Yu, K. F., & Levine, R. J. (2007). Misclassification of maternal smoking status and its effects on an epidemiologic study of pregnancy outcomes. *Nicotine & Tobacco Research*, *9*, 1005–1013.

Espy, K. A., Fang, H., Johnson, C., Stopp, C., Wiebe, S., & Respass, J. (in press). Prenatal tobacco exposure: Developmental impact on neonatal regulation. *Developmental Psychology*.

Fergusson, D., Goodwin, R., & Horwood, R. (2003). Major depression and cigarette smoking: Results of a 21-year longitudinal study. *Psychological Medicine*, *33*, 1357–1367.

Flick, L., Cook, C., Homan, S., McSweeney, M., Campbell, C., & Parnell, L. (2006). Persistent tobacco use among during pregnancy and the likelihood of psychiatric disorders. *American Journal of Public Health*, *96*, 1799–1807.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*, 367–378.

Goodwin, R., Keyes, K., & Simuro, N. (2007). Mental disorders and nicotine dependence among pregnant women in the United States. *Obstetrics and Gynecology*, *109*, 875–883.

Imbens, G. (2003). *Nonparametric estimation of average treatment effects under exogeneity: A review*, (National Bureau of Economic Research, Technical Report, T0294) Retrieved from http://www.nber.org/papers/t0294

Kodl, Middlecamp M., & Wakschlag, L. (2004). Does a childhood history of externalizing problems predict smoking during pregnancy? *Addictive Behaviors*, *29*, 273–279.

Lumley, J. (1987). Stopping smoking. *British Journal of Obstetrics & Gynecology*, *94*, 289–294.

McCaffrey, D., Ridgeway, G., & Morral, A. (2004). Propensity score estimation with boosted regression for evaluating adolescent substance abuse treatment. *Psychological Methods*, *9*, 403–425.

Muthén, L., & Muthén, B. (2007). *Mplus user's guide* (5th ed.). Los Angeles, CA: Author.

Parton, G., Carlin, J., Coffey, C., Wolfe, R., Hibbert, M., & Bowes, G. (1998). Depression, anxiety, and smoking initiation: A prospective study over 3 years. *American Journal of Public Health*, *88*, 1518–1522.

Pickett, K., Wilkinson, R., & Wakschlag, L. (2009). The psychosocial context of pregnancy smoking and quitting in the Millennium Cohort Study. *Journal of Epidemiology and Community Health*, *63*, 474–480.

Pley, E., Wouters, E., Voorhorst, F., Stolte, S., Kurver, P., & de Jong, P. (1991). Assessment of tobacco-exposure during pregnancy; behavioural and biochemical changes. *European Journal of Obstetrics, Gynecology and Reproductive Biology*, *40*, 197–201.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0 Retrieved from http://www.R-project.org

Ridgeway, G. (2006). Assessing the effect of race bias in post-traffic stop outcomes using propensity scores. *Journal of Quantitative Criminology*, *22*, 1–29.

Riese, M. (1982). Procedures and norms for assessing behavioral patterns in full-term and stable pre-term neonates. *JSAS Catalog of Selected Documents in Psychology*, *12*(MS. No.2415).

Riese, M. (1986). Implications of sex differences in neonatal temperament for early risk and developmental/environmental interactions. *Journal of Genetic Psychology*, *147*, 507–513.

Rodriguez, A., Bohlin, G., & Lindmark, G. (2000). Psychosocial predictors of smoking and exercise during pregnancy. *Journal of Reproductive and Infant Psychology*, *18*, 203–223.

Rosenbaum, P. (2002). *Observational studies*. New York, NY: Springer.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.

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

Satorra, A. (2000). Scaled and adjusted restricted tests in multisample analysis of moment structures. In R. D. H. Heijmans, D. S. G. Pollock & A. Satorra (Eds.), *Innovations in multivariate statistical analysis. A Festschrift for Heinz Neudecker* (pp. 233–247), London: Kluwer Academic.

Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictors. *Machine Learning*, *37*, 297–336.

Schuetze, P., & Eiden, R. (2006). The association between maternal smoking and secondhand exposure and autonomic functioning at 2–4 weeks of age. *Infant Behavior and Development*, *29*, 32–43.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton-Mifflin.

Shiono, P. H., Klebanoff, M., & Rhoads, C. (1986). Smoking and drinking during pregnancy. *Journal of the American Medical Association*, *255*, 82–84.

Wang, J., & Donnan, P. T. (2001). Propensity score methods in drug safety studies: Practice, strengths and limitations. *Pharmacoepidemiology and Drug Safety*, *10*, 341–344.

West, S. G., Biesanz, J. C., & Pitts, S. C. (2000). Causal inference and generalization in field settings experimental and quasiexperimental designs. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 40–88), New York: Cambridge University Press.

Whiteman, M., Fowkes, F., Deary, I., & Lee, A. (1997). Hostility, cigarette smoking and alcohol consumption in the general population. *Social Science Medicine*, *44*, 1089–1096.

Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities: Brief Intellectual Ability*. Itasca, IL: Riverside Publishing.

Yolton, K., Khoury, J., Xu, Y., Succop, P., Lanphear, B., Bernert, J. T., et al. (2009). Low-level prenatal exposure to nicotine and infant neurobehavior. *Neurotoxicology and Teratology*, *31*, 356–363.

Yuan, K., & Bentler, P. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. E. Sobel & M. P. Becker (Eds.), *Sociological methodology 2000* (pp. 165–200), Washington, DC: ASA.