HSR Health Services Research

DOI: 10.1111/1475-6773.12068
METHODS ARTICLE

An Empirical Comparison of Tree-Based Methods for Propensity Score Estimation

Stephanie Watkins, Michele Jonsson-Funk, M. Alan Brookhart, Steven A. Rosenberg, T. Michael O'Shea, and Julie Daniels

Objective. To illustrate the use of ensemble tree-based methods (random forest classification [RFC] and bagging) for propensity score estimation and to compare these methods with logistic regression, in the context of evaluating the effect of physical and occupational therapy on preschool motor ability among very low birth weight (VLBW) children.

Data Source. We used secondary data from the Early Childhood Longitudinal Study Birth Cohort (ECLS-B) between 2001 and 2006.

Study Design. We estimated the predicted probability of treatment using tree-based methods and logistic regression (LR). We then modeled the exposure-outcome relation using weighted LR models while considering covariate balance and precision for each propensity score estimation method.

Principal Findings. Among approximately 500 VLBW children, therapy receipt was associated with moderately improved preschool motor ability. Overall, ensemble methods produced the best covariate balance (Mean Squared Difference: 0.03–0.07) and the most precise effect estimates compared to LR (Mean Squared Difference: 0.11). The overall magnitude of the effect estimates was similar between RFC and LR estimation methods.

Conclusion. Propensity score estimation using RFC and bagging produced better covariate balance with increased precision compared to LR. Ensemble methods are a useful alterative to logistic regression to control confounding in observational studies.

Key Words. Propensity scores, tree-based methods, ensemble methods

Evaluation of treatment effectiveness in nonrandomized studies is complicated by exposure group differences on measured and unmeasured characteristics that are independently related to the outcome. The resulting confounding can be controlled using propensity scores to balance observed confounders between treatment groups. The propensity score represents the probability of receiving treatment conditional on a set of confounders

(Rosenbaum and Rubin 1983). Individuals with similar propensity scores can be expected to have similar values on measured background characteristics. Given a correctly specified propensity score model, once one conditions on the propensity score, differences in measured characteristics between the treatment groups should be from chance alone, assuming no unmeasured confounders (Rosenbaum and Rubin 1983).

The true propensity score, or the predicted probability of treatment, is not known. Therefore, researchers must estimate the propensity score, typically using parametric models (Austin 2011). Logistic regression is one of the most common methods used to estimate propensity scores, but it requires several assumptions. The relation between continuous and ordinal independent variables and the log odds of the dependent variable must be linear. Furthermore, the model assumes additivity. Therefore, researchers must consider the functional form of covariates as well as interaction terms (D'Agostino 1998). Violations can result in misspecification of the propensity score model and a biased effect estimate (Drake 1993).

Regression tree-based methods, including bagging and random forest classification (RFC), are nonparametric methods derived from learning-based algorithms that offer alternative strategies for estimating the propensity score. The methods use a series of classification trees to estimate the average probability of membership in a given class. These techniques may improve predictive accuracy compared to classical statistical techniques such as linear and logistic regression (Breiman 2001b). For example, in simulation studies, regardless of nonlinearity or nonadditivity, random forest performed well in terms of covariate balance between treatment groups and may result in further reduction in bias of the effect estimate when compared to traditional logistic regression (Setoguchi et al. 2008; Lee, Lessler, and Stuart 2010).

There has been relatively little investigation into the use of tree-based methods to estimate the propensity score (Westreich, Lessler, and Funk 2010). In this article, we illustrate the use of three tree-based methods: bagging, RFC,

---

Address correspondence to Stephanie Watkins, Ph.D., M.S.P.H., M.S.P.T., Center for Health Promotion and Disease Prevention, University of North Carolina at Chapel Hill, University of North Carolina, 1700 Martin Luther King Blvd CB#7426, Chapel Hill, NC 27599-7426; e-mail: wat@email.unc.edu. Michele Jonsson-Funk, Ph.D., M. Alan Brookhart, Ph.D., and Julie Daniels, Ph.D., are with the Department of Epidemiology, UNC Gillings School of Global Public Health, Chapel Hill, NC. Steven A. Rosenberg, Ph.D., is with the Department of Psychiatry, University of Colorado at Denver and Health Sciences Center, Denver, CO. T. Michael O'Shea, M.D., M.P.H., is with the Department of Pediatrics, Wake Forest University School of Medicine, Winston-Salem, NC.

and a single classification tree. We evaluate these methods in the context of an analysis to understand the effect of physical and occupational therapy services on the motor skills of preschoolers who were born with very low birth weight (VLBW). We consider the propensity scores estimated by tree-based methods in comparison to a logistic regression model. We then compare the distribution of the estimated propensity scores, the balance of covariates, and the change in effect estimates after applying inverse probability of treatment weights (IPTW).

Conceptual Overview

Classification Trees

Classification tree analysis is a nonparametric method commonly used in data mining where a set of independent variables are used to predict membership of observations in a given class of the dependent variable. The method evaluates the relation between predictors and treatment with a learning algorithm using decision trees to partition observations into nodes with similar probabilities of class membership in the treatment group (Breiman 2001b). The dataset is partitioned until nodes, or branches of the tree, are as homogenous as possible with respect to class membership (Breiman et al. 1984). The tree begins with a root node and continues to split until the nodes reach either a given sample size or a given level of impurity reduction. With each partition of the tree, groups of subjects with the majority of a given level of the response are isolated. This daughter node is considered to have a more "pure" or homogenous response compared to the parent node. Researchers may set a stopping rule, allowing the classification tree to stop splitting once the difference in impurity between the parent and daughter nodes reaches a given threshold. At each terminal node, the algorithm predicts the response class by classifying the response according to the class that received the largest number of votes (Strobl, Malloy, and Tutz 2009).

Despite the popularity of this data mining method, results from a single classification tree are highly variable and are known to be unstable (Strobl, Malloy, and Tutz 2009). For example, the rank of each variable in the classification tree as well as the cut point of the variable is strongly dependent upon the distribution of observations in the data. With small changes in the data structure, the order of variable selection or the cut point of the variable may change, resulting in an alternative tree structure (Strobl, Malloy, and Tutz 2009).

Bagging and Random Forest Classification

Both bagging and RFC are tree-based methods that attempt to improve the stability of tree-based methods that rely on a single tree. These methods aggregate predictions over multiple individual classification trees to improve the overall predictive performance of the algorithm. Bagging, or bootstrap aggregation, randomly draws a series of bootstrap samples from the data and creates individual classification trees for each sample. With each bootstrap sample, the data will vary slightly from the previous sample. Furthermore, each individual tree can vary, perhaps substantially, from the previous tree. The algorithm then aggregates the predicted probability of class membership over the series of classification trees (Breiman 2001a).

Random forest classification also utilizes the same bootstrap aggregation approach. However, random forest adds an additional level of variability to the algorithm. During construction of the individual classification trees, a random sample of predictor variables is chosen to split the data at each node. Therefore, each individual tree is even more diverse compared to the trees from bagging alone (Strobl, Malloy, and Tutz 2009).

Although individual classification trees are inherently unstable, bagging and RFC have been shown to produce robust estimates. In both empirical and simulation studies, estimates aggregated over a series of classification trees show improvements in prediction accuracy when compared to a single classification tree (Breiman 1996, 1998; Dietterich 2000; Buhlmann and Yu 2002). Bagging may equalize the influence of given observations in the data (Strobl, Malloy, and Tutz 2009). Thus, data points that strongly influence the classification algorithm are downweighted (Strobl, Malloy, and Tutz 2009). Furthermore, the additional level of variability introduced by RFC creates additional diversity between trees with a smaller upper bound of error (Breiman 2001a). Overall, these methods produce a more robust final estimate with decreased variability (Breiman 2001a).

METHODS

We illustrate the use of three tree-based methods: bagging, RFC, and a single classification tree, as well as parametric logistic regression in an analysis that evaluates the effect of physical and occupational therapy services on motor performance among preschool children who are VLBW and "at risk" for developmental coordination disorder (DCD). DCD is a condition defined as

an impairment in the development of motor coordination among children without known physical or neurological impairments (American Psychiatric Association 2000). Children with DCD are at increased risk for low academic performance, low self-esteem, and limited physical activity which may continue into adolescence (Cairney et al. 2005a,b; Missiuna et al. 2007). Children who are born VLBW are six times as likely to have DCD compared to their normal birth weight peers (Edwards et al. 2011).

Population and Variables

This study is described in detail elsewhere (Watkins et al. unpublished data). Briefly, using data from the Early Childhood Longitudinal Study Birth Cohort (ECLS-B), our sample included approximately 500 VLBW children who were without known mobility problems and appeared to be meeting normal developmental motor milestones at 9 months. Researchers asked families between 9 months and 2 years of age whether their child had ever received physical or occupational therapy services. We considered the child exposed if the child ever received either therapy during this time period.

Researchers directly assessed preschool gross motor performance using items from the Bruininks-Oseretsky Test of Motor Proficiency (Bruininks, 1978) the Movement Assessment Battery for Children (Henderson and Sugden 1992), and the Early Screening Inventory-Revised (Meisels et al. 1997). These norm referenced assessments are commonly used to evaluate motor ability among preschool children (Rydz et al. 2005; Piek, Gasson, and Summers 2008). Researchers directly reported the child's ability to complete the following locomotor tasks on a pass/fail basis: skipping eight consecutive steps, hopping on one foot five times, walking backwards six steps on a taped line, and standing on one foot for 10 seconds.

On the basis of a priori substantive knowledge, we created a directed acyclic graph and determined a minimum sufficient conditioning set of confounders (VanderWeele and Robins 2009). A directed acyclic graph is a diagram that provides a graphical representation of the causal relation between two variables. The diagram guides the researcher in determining confounders of the relation between the treatment and the outcome (Rothman, Greenland, and Lash 2008). The final confounders in our analysis included the following: gestational age, birth weight, length of the child's hospital stay after birth, age at which the child began to walk with assistance, race/ethnicity, parental education, socioeconomic status, and the child's 9-month Bayley Short Form-Research Edition (BSF-R) motor T score. The BSF-R is a subset of

items taken from the standardized Bayley Scales of Infant Development, Second Edition to assess children's cognitive, motor, and language skills (Bayley 1993).

Propensity Score

In this observational study, treatment assignment into physical and occupational therapy services was not randomized. Children and their families may participate in therapy treatment based on a host of factors including their child's functional ability and access to health care. Therefore, the distribution of baseline characteristics between children in the treated and untreated groups may differ and children between these two groups would not be "exchangeable". We estimated the average treatment effect of early childhood physical and occupational therapy using a propensity score approach to control for confounding.

Estimating the Propensity Score

We estimated the conditional probability of treatment given the identified confounders stated above using the following four methods: bagging, RFC, a single classification tree, and logistic regression.

Using the R statistical platform (Gentleman and Ihaka 2008), we first used the RandomForest package (Liaw and Wiener 2002) to estimate the predicted probability of class membership in the therapy group given the following covariates: gestational age, birth weight, length of the child's hospital stay after birth, age at which the child began to walk with assistance, race/ethnicity, parental education, socioeconomic status, and the child's 9-month BSF-R Motor T score. Race/ethnicity and parental education were entered as a series of indicator variables; all other variables were entered as continuous variables. We set the random forest algorithm to generate 1,000 individual classification trees. The suggested default for the number of random splitting variables at each node is the square root of the number of variables in the algorithm (Liaw and Wiener 2002). Our model included 19 variables, so we set the default to four variables chosen at each split.

We checked the sensitivity of the error rate to our chosen parameters by allowing the number of trees to vary between 250 and 1,000 and the number of randomly chosen variables to vary between two and seven. The error rate for the algorithm is generated from the 33 percent of the data remaining that were not used to form the classification trees. For example, with each

bootstrap sample, the remaining data (≈33 percent) not in the sample are entered into the classification tree. The error in these out-of-bag predictions is collected over the series of trees to determine the final error rate over the forest. The error rate is considered to be robust if the predicted probabilities of class membership are aggregated across a sufficient number of trees (Liaw and Wiener 2002). However, if the number of trees are too few, then the error rate may be upwardly biased (Bylander 2002). The algorithm may, therefore, be a better predictor of the outcome than suggested by the error rate.

We then implemented the Ipred package (Peters and Hothorn 2012) and the Tree package (Ripley 2012) using the R statistical software to estimate the predicted probabilities of having class membership in the treatment group using bagging and a single classification tree, respectively. For both models, we entered the same covariates as in the RFC algorithm. In the Ipred package, we generated a series of 1,000 trees and checked the sensitivity of the error rate by varying the number of trees between 250 and 1,000. For both methods, the splitting variables were chosen by the algorithm in a hierarchical fashion based on impurity reduction.

Lastly, we generated predicted probabilities of receiving physical or occupational therapy using logistic regression. As in common practice, we entered potential confounders as main effects. Race/ethnicity and parental education were modeled as indicator variables; all others were entered into the model as continuous terms. In addition, we considered the functional form of the covariates, of which gestational age and baseline motor ability appeared to have a U-shaped association with preschool motor ability. We entered these two covariates as quadratic terms.

Statistical Analysis

We generated unique inverse probability of treatment weights using each method: RFC, bagging, a single classification tree, and logistic regression. These weights create a pseudo population of children with a distribution of covariates that represents the combined sample (Bang and Robins 2005). To estimate the average treatment effect, treated children received a weight of (1/propensity score). Children in the untreated group received a weight of (1/(1-propensity score)). To evaluate the balance of each propensity score method, we then calculated the standardized difference of the weighted confounding variables between the treatment groups.

Standardized differences represent the differences between the means by therapy status in units of standard deviations. The estimate is calculated as

$$d = |\bar{x}_{\text{therapy}} - \bar{x}_{\text{notherapy}}| \sqrt{s^2_{\text{therapy}} + s^2_{\text{notherapy}}/2}$$

(Flury and Riedwyl 1986). Although there is no standard criterion to determine balance between treatment groups, experts suggest a standardized difference of <0.10 (Normand et al. 2001; Austin and Mamdani 2006; Austin 2007). We then averaged the standardized differences across all confounders to determine the mean standardized difference (MSD).

Finally, for each of the four methods, we estimated the average effect of physical and occupational therapy on preschool motor performance using logistic regression and IPTW in SAS version 9.2 (SAS Institute, Inc., Cary, NC, USA). In practice, inverse probability of treatment weights may often be highly variable. For example, weights may be extreme for treated subjects with a low propensity for treatment or for untreated subject with a high propensity for treatment. Stabilization of the weights is suggested to decrease the variance, which provides a narrower confidence interval around the estimated effect estimate. We stabilized the weights by multiplying the child's IPTW by the marginal probability of the treatment that he or she actually received (Cole and Hernan 2008).

Missing Data

In these data, approximately 7 percent of children were missing at least one covariate used to estimate the propensity score. Data on motor ability was missing for approximately 20 percent of the sample. We included only children with complete data to estimate the predicted probability of treatment. Thus, we compare the balance of covariates and the estimated effect estimates for each method among the same group of children.

RESULTS

A description of this cohort has been presented elsewhere (Watkins et al. unpublished data). Briefly, the sample included approximately 500[1] children weighing less than 1,500 g at birth of which 6.5 percent of children received therapy between 9 months and age 2 years. Children who received therapy were more likely to be white (58.1 percent vs. 38.5 percent) and of male gender (61.3 percent vs. 45.3 percent) and were born 2 weeks earlier in gestation, on average. Developmentally, the treated children sat independently, crawled, and walked with assistance on average, 1 month later than the

untreated children. Five-minute APGAR scores were similar between the two groups.

Random Forest Classification/Bagging: Error Rate

In our sample of approximately 450 children with complete covariate data, the random forest algorithm misclassified treatment status 15.7 percent of the time over 1,000 trees with four variables randomly chosen at each split. Overall, there was little change in the error rate with small changes in the number of splitting variables. The error rate over our chosen range of trees and number of splitting variables varied by approximately 0.5 percent. The misclassification rate for the bagging algorithm over 1,000 trees was 15.7 percent. The misclassification rate increased to 17.0 percent with only 250 trees (Table 1).

Propensity Score

The mean predicted probability of receiving treatment for the children who received therapy ranged between 0.16 and 0.20 across the RFC and bagging tree-based methods and the main effects logistic regression model. The single classification tree yielded a predicted probability of treatment that was approximately twice that of the other three methods for children who received therapy. The mean predicted probability of treatment for children who did not receive therapy ranged between 0.05 and 0.07 across all four methods used to estimate the propensity score. Children in the treatment groups received similar weights across estimation methods with the exception of the single classification tree algorithm, which led to a higher propensity for

Table 1: Out-of-Bag Error Rates* for Prediction of Receipt of Early Childhood Therapy

	Out-of-Bag Error		
Number of trees	250	750	1,000
Random forest	%	%	%
Number of randomly chosen variables per split			
2	15.4	15.4	15.4
4	15.4	15.4	15.7
7	15.9	16.1	16.1
Bagging	17.0	16.6	15.7

*The out-of-bag data are put down each bootstrap classification tree, and the results are aggregated to determine the out-of-bag error rate over the forest of trees.

Table 2: Distribution of Propensity Score and Weights for the Average Treatment Effect by Method Used to Estimate the Propensity Score*,†

	Random Forest‡ Classification		*Logistic Regression*		*Classification Tree*		*Bagging‡*	
	Early Childhood Therapy	*No Early Childhood Therapy*	*Early Childhood Therapy*	*No Early Childhood Therapy*	*Early Childhood Therapy*	*No Early Childhood Therapy*	*Early Childhood Therapy*	*No Early Childhood Therapy*
Propensity score§								
Minimum	0.01	0.00	0.02	0.00	0.06	0.00	0.01	0.00
Maximum	0.42	0.54	0.43	0.56	0.64	0.64	0.55	0.59
Mean	0.16	0.06	0.16	0.07	0.39	0.05	0.20	0.07
Average treatment effect weights								
Minimum	0.16	0.93	0.15	0.94	0.10	0.93	0.12	0.93
Maximum	4.91	2.01	2.81	2.15	1.08	2.57	8.00	2.26
Mean	0.89	1.00	0.79	1.01	0.27	1.00	0.99	1.01

*Early Childhood Longitudinal Study, Birth Cohort 2001–2006.
†Average treatment effect weight is estimated as (1/propensity score) for those children who received early childhood therapy. For the those children who did not receive therapy, the weight is (1/(1−propensity score)). These weights are stabilized so the sum of the weights reflects the size of the original population. We multiplied the weight by the probability of receiving the treatment that the child actually received.
‡We estimated the propensity score using random forest classification. The out-of-bag error rate for the algorithm for classification of receipt of early childhood therapy was 15.7% across 1,000 trees, where the algorithm chose four random variables at each split of the node. The error rate for bagging was also 15.7% over 1,000 trees.
§The propensity score includes the following confounders: 9 months BSF-R motor T score, socioeconomic status, length of child's hospital stay after birth, gestational age, birth weight, parental education, race, and age at which the child walked with assistance.

treatment and a lower weight compared to the other estimation methods. The weights for children who did not receive physical or occupational therapy were similar for all four methods (Table 2).

Covariate Balance

In the unweighted sample, the MSD across strong confounding covariates was 0.54. The length of the infant's hospital stay after birth (Standardized Difference: 0.73) and the age at which the child crawled and the child's gestational age (Standardized Difference: 0.74 and 0.65 respectively) were most unbalanced between the treatment groups. After applying the weights estimated by the RF and bagging tree-based methods, the distribution of baseline covariates differed only negligibly by treatment status. The MSD across covariates was 0.07 using the random forest method and 0.03 using the bagging algorithm. After implementing the random forest algorithm, length of hospital stay and gestational age remained slightly unbalanced (standardized difference: 0.14 and 0.15, respectively). The mean length of hospital stay and gestational age after applying the bagging method was quite similar (standardized difference: 0.03 and 0.07, respectively) by therapy status (Table 3).

The MSD for the covariates weighted with the logistic model was 0.11. The standardized difference for birth weight, length of hospital stay, and age at crawling and walking with assistance was greater than the suggested 0.10 criterion for these covariates. When the propensity score was estimated by the single classification tree covariates differed by approximately 0.18 standard deviations (Table 3).

Multivariable Regression

In the weighted multivariable logistic regression models, receipt of interventional physical or occupational therapy services between 9 months and age 2 years was moderately associated with improvement in preschool coordination skills in this VLBW population. However, overall this association did not reach statistical significance. The effect was consistent across both the tree-based methods as well as the logistic method used to estimate the propensity score; however, the magnitude of the effect as well as the precision of the estimate varied by method. The random forest algorithm produced the most precise estimate in the weighted model for hopping and single leg stance. The bagging algorithm produced slightly more precise estimates for

Table 3: Standardized Differences among Confounders by Propensity Score Method

	Unweighted			*Random Forest Classification*		
Confounder	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**
Birth weight (g)	1,030.23 (212.63)	1,147.77 (250.31)	0.51	1,095.65 (746.04)	1,139.93 (259.76)	0.08
BSF-R motor T score	45.16 (10.85)	49.81 (8.44)	0.48	49.00 (32.91)	49.44 (8.88)	0.02
Days in hospital after birth	67.94 (22.33)	48.46 (30.40)	0.73	58.30 (87.00)	49.27(31.78)	0.14
SES	−0.14 (0.71)	−0.27 (0.80)	0.17	−0.27 (2.66)	−0.25 (0.84)	0.01
Age of independent sitting (months)	8.45 (1.57)	7.77 (1.73)	0.41	7.83 (6.53)	7.84 (1.82)	0.00
Age of crawling (months)	9.81 (1.63)	8.56 (1.74)	0.74	9.12 (7.62)	8.62 (1.79)	0.09
Age of walking with assistance (months)	11.00 (1.53)	10.00 (1.59)	0.64	10.26 (4.96)	10.07 (1.66)	0.05
Gestational age (weeks)	27.84 (3.17)	30.13 (3.80)	0.65	28.78 (11.16)	30.04 (3.92)	0.15
Mean			0.54			0.07
	Logistic Regression			*Classification Tree*		
Confounder	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**
Birth weight (g)	1,047.03 (673.98)	1,138.16 (262.88)	0.18	1,054.48 (418.10)	1,142.43 (257.16)	0.25
BSF-R motor T score	48.34 (34.29)	49.36 (8.85)	0.04	47.79 (17.99)	49.33 (8.95)	0.11
Days in hospital after birth	60.97 (66.65)	49.60 (32.32)	0.22	59.41 (47.86)	49.61 (31.50)	0.24
SES	−0.26 (2.46)	−0.25 (0.85)	0.00	−0.11 (1.29)	−0.23 (0.85)	0.12
Age of independent sitting (months)	8.11 (5.76)	7.85 (1.83)	0.06	7.56 (3.52)	7.83 (1.80)	0.10
Age of crawling (months)	9.45 (5.95)	8.63 (1.81)	0.19	9.22 (4.01)	8.61 (1.78)	0.20

continued

Table 3. *Continued*

	Logistic Regression			*Classification Tree*		
Confounder	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**	*Early Childhood Therapy Mean (SE)*	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**
Age of walking with assistance (months)	10.44 (4.42)	10.08 (1.67)	0.11	10.38 (2.89)	10.07 (1.65)	0.13
Gestation age (weeks)	29.28 (13.57)	29.99 (3.93)	0.07	28.45 (6.11)	30.04 (3.90)	0.31
Mean			0.11			0.18

	Bagging		
	Early Childhood Therapy Mean (SE)	*No Early Childhood Therapy Mean (SE)*	*Standardized Differences**
Birth weight (g)	1,146.48 (673.17)	1,138.75 (262.42)	0.02
BSF-R motor Tscore	48.53 (31.48)	49.35 (9.00)	0.04
Days in hospital after birth	51.26 (85.37)	49.58 (31.97)	0.03
SES	−0.25 (2.92)	−0.24 (0.85)	0.00
Age of independent sitting (months)	7.55 (7.04)	7.85 (1.84)	0.06
Age of crawling (months)	8.63 (8.22)	8.64 (1.82)	0.00
Age of cruising (months)	10.10 (5.19)	10.09 (1.68)	0.00
Gestational age (weeks)	29.43 (11.17)	30.02 (3.95)	0.07
Mean			0.03

*Standardized differences represent the differences between the means by therapy status in units of standard deviations. The estimates are calculated as

$$d = |\bar{x}_{\text{therapy}} - \bar{x}_{\text{notherapy}}| \sqrt{s^2_{\text{therapy}} + s^2_{\text{notherapy}}/2}$$

(Flury and Riedwyl 1986).
SES, socioeconomic status.

Table 4: Average Treatment Effect of Interventional Physical or Occupational Therapy Services and Preschool Motor Skills: Using Three Methods to Generate the Propensity for Treatment*,‡,§

Preschool Motor Outcomes	*Crude Model*				*Random Forest Classification*†				*Logistic Regression Main Effects*¶			*Logistic Regression Quadratic Terms*‖			*Bagging*†		
Preschool Motor Skills†	*N*	*OR*	*95% CI*	*CLR*	*N*	*OR*	*95% CI*	*CLR*	*OR*	*95% CI*	*CLR*	*OR*	*95% CI*	*CLR*	*OR*	*95% CI*	*CLR*
Skipping eight consecutive steps†	300	0.85	0.28, 2.63	9.50	300	2.10	0.71, 6.17	8.64	2.01	0.65, 6.23	9.58	2.05	0.65, 6.47	9.92	3.06	1.04, 8.98	8.64
Hopping five times independently†	300	0.56	0.23, 1.32	5.69	300	0.75	0.28, 2.01	7.20	0.82	0.29, 2.27	7.74	0.95	0.34, 2.69	7.99	1.02	0.37, 2.84	7.78
Maintaining single leg stance for 10 seconds independently†	350	0.74	0.32, 1.71	5.38	350	1.07	0.42, 2.84	6.74	0.87	0.32, 2.34	7.20	0.95	0.35, 2.53	7.15	1.70	0.63, 4.58	7.24
Walking backward six steps on a line†	350	1.01	0.39, 2.63	6.82	350	1.45	0.49, 4.25	8.62	1.46	0.48, 4.38	9.06	1.39	0.45, 4.32	9.54	1.88	0.86, 7.06	8.20

*Early Childhood Longitudinal Study, Birth Cohort 2001–2006. Counts rounded to the nearest 50 according to data use agreement. These data were weighted using (1/propensity score) for children who received therapy and (1/(1—propensity score)) for those children who did not receive treatment. The propensity scores were stabilized by multiplying the treatment weights by the marginal prevalence of the treatment that they actually received.

†The out-of-bag error rate for the algorithm for classification of receipt of early childhood therapy was 15.7% across 1,000 trees, where the algorithm chose four random variables at each split of the node. The out-of-bag error rate was also 15.7% for the bagging algorithm across 1,000 trees.

‡We defined children who received physical and occupational therapy between 9 months and age 2 years as treated.

§Overall, 7% percent of children were missing at least one covariate used to estimate the propensity score and were excluded. Outcome data were missing for approximately 20% of the sample, and 6% of children were missing data on receipt of therapy.

¶Includes only main effect terms for confounders in the propensity score model: 9-month BSF-R motor T score, socioeconomic status, length of child's hospital stay after birth, gestational age, birth weight, parental education, race, and age at which the child walked with assistance.

‖9-month BSF-R motor T score and gestation age entered as quadratic terms.

the walking backward task (Table 4). When we used logistic regression to estimate the propensity score, with either main effects or quadratic terms, overall the confidence intervals for the effect estimates were the least precise.

The magnitude of the estimate for skipping ability was largest (OR: 3.06, 95 percent CI: 1.04, 8.98) using the bagging technique and smallest (OR: 2.01, 95 percent CI: 0.65, 6.23) using logistic regression with main effects to estimate the propensity score. The bagging estimate continued to generate the effect estimates of the greatest magnitude for the additional motor outcomes modeled in these data. The magnitude of the effect estimates produced with the logistic regression main effects model and the model that included quadratic terms were similar. In general, logistic regression estimation of the propensity score produced the most conservative effect estimates for the majority of preschool motor items. The single classification tree algorithm did not balance the covariates well between treatment groups, and therefore the results of the weighted models using this method are not presented (Table 4).

DISCUSSION

In this article, we illustrate the use of various tree-based methods to estimate the predicted probability of receiving interventional physical and occupational therapy services in a sample of VLBW children. Furthermore, we considered how propensity scores estimated from bagging and RFC balanced covariates between treatment groups and compared these methods with the performance of propensity scores estimated from a single classification tree as well as traditional logistic regression.

In our sample, bagging and RFC achieved the best overall balance of covariates across treatment groups. Among all methods used to estimate the propensity score, the mean standardized difference of all covariates was smallest for these two methods. The propensity scores estimated from the logistic model showed a marginal imbalance in covariates, where the single classification tree method had the worst performance.

These findings are supported by the study of Lee and colleagues who studied machine learning methods when estimating the propensity score in simulated data. In a small sample, when compared to standard logistic regression and a single classification tree, random forest and bagging returned the lowest mean absolute standardized differences. The standardized differences

between individual covariates were also less dispersed with these two methods. The resulting bias in these simulated models was highest when the propensity score was estimated from a single classification tree and lowest when the propensity score was estimated using either bagging (10.3 percent) or RFC (7.7 percent) (Lee, Lessler, and Stuart 2010).

In our data, it appeared that propensity score estimation using logistic regression did a reasonable job of balancing the covariates between our treatment groups. However, it is not known how well this model performed in reducing the amount of bias because the true treatment effect is not known. While the effect estimates assessing preschool coordination were similar between the two models when we estimated the propensity score by RFC and main effects logistic regression, the effect estimates for the child's ability to balance differed by approximately 19 percent. However, when we included a quadratic term in our propensity score model, the effect estimates differed by only 12 percent.

In simulation studies, a main effects logistic regression model performed adequately in reducing bias when the relation between independent variables and the logit of the outcome is linear and additive (Lee, Lessler, and Stuart 2010). However, researchers reported a mean absolute bias of 30 percent in the presence of nonlinearity and nonadditivity (Lee, Lessler, and Stuart 2010). For comparison, we used a main effects logistic regression model that appears to be commonly used by researchers as well as a logistic regression model where we included quadratic terms. In our data, the relation between several confounders and the logit of receiving treatment was curvilinear. Due to our small sample size, we were limited in our ability to test for interactions. For the balancing task, our effect estimate weighted with the propensity score estimated from the logistic regression model with quadratic terms more closely approximated our effect estimates weighted with propensity scores from RFC. Therefore, the difference in the estimated effects between the main effects logistic regression model and RFC may be due to lack of consideration of the relation between confounders and the logit of receiving treatment. By modeling the functional form of the variable, and considering interactions, the logistic regression model may be more effective. However, the nonparametric random forest algorithm is inherently flexible for incorporating interactions as well as nonlinear functional forms which may be more feasible in some circumstances.

In our analysis, ensemble tree-based methods, including random forest and bagging, appear to outperform traditional logistic regression methods with main effects. In our analysis, both tree-based methods performed well in

balancing the covariates between treatment groups; however, the bagging method resulted in effect estimates of greater magnitude. It is possible that the additional level of randomness implemented by the random forest classifier allowed less important variables to be expressed in predicting therapy exposure, thereby attenuating the magnitude of the effects.

Propensity scores are a useful tool to control for confounding in children's health research. Yet the method is subject to several limitations. Propensity scores only control for measured confounders in the data. Therefore, residual confounding in the estimated effect estimate may still be present due to unmeasured confounders. Moreover, without a careful modeling technique, one may mispecify the propensity score model and the estimated treatment effect may be biased.

In this study, estimation of the propensity score using ensemble tree-based methods produced the smallest standardized differences across covariates. The resulting effect estimates varied slightly depending on the method used to estimate the propensity score. Although we are unsure of the true effect estimate, studies show that the effect estimates derived from RFC and bagging are the least biased and logistic regression may adequately reduce bias in the presence of additivity and linearity (Lee, Lessler, and Stuart 2010).

Estimation of the propensity score using tree-based ensemble methods may be a useful method to evaluate the effect of interventions on childhood motor skills while controlling for confounding. These methods appear to be robust, creating better covariate balance for control of confounding and a potential for further bias reduction compared to main effects logistic regression.

In our study, using ensemble tree-based methods to adjust for confounding, we found that early intervention physical and occupational therapy services were moderately beneficial for select preschool motor skills. To date, few studies have examined the impact of these services on preschool motor ability among VLBW children. However, interventions that promote motor skills among children of normal birth weight do appear to benefit object control and locomotion in early childhood (Riethmuller, Jones, and Okely 2009). Our findings, although not statistically significant, support the delivery of early-intervention physical and occupational therapy services to VLBW children who are at risk for poor motor coordination. However, future work is needed both to confirm our conclusions about efficacy as well as to examine the influence on efficacy of frequency and duration of therapy.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: This project was supported, in part, by grant number Ko2hs01 7950 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality. Dr. Jonsson Funk has received salary support from the Center for Pharmacoepidemiology, which is funded by an unrestricted grant from Glaxsmithkline.

Disclosures: None.

Disclaimers: None.

NOTE

1. Numbers are rounded to the nearest 50 to protect the privacy of participants per the ECLS-B data use agreement.

REFERENCES

American Psychiatric Association 2000. *Diagnostic and Statistical Manual of Mental Disorders.* Washington, DC: American Psychiatric Association.

Austin, P. C. 2007. "Propensity-Score Matching in the Cardiovascular Surgery Literature from 2004 to 2006: A Systematic Review and Suggestions for Improvement." *Journal of Thoracic and Cardiovascular Surgery* 134 (5): 1128–35.

——— .2011. "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies." *Multivariate Behavioral Research* 46 (3): 399–424.

Austin, P. C., and M. M. Mamdani. 2006. "A Comparison of Propensity Score Methods: A Case-Study Estimating the Effectiveness of Post-AMI Statin Use." *Statistics in Medicine* 25 (12): 2084–106.

Bang, H., and J. M. Robins. 2005. "Doubly Robust Estimation in Missing Data and Causal Inference Models." *Biometrics* 61 (4): 962–73.

Bayley, N.. 1993. *Bayley Scales of Infant Development. Motor Scale Record Form*, 2nd Edition. San Antonio, TX: Psychological Corporation.

Breiman, L. 1996. "Bagging Predictors." *Machine Learning* 24: 123–40.

——— .1998. "Arcing Classifiers." *Annals of Statistics* 26: 801–49.

——— .2001a. "Random Forests." *Machine Learning* 45: 5–32.

——— .2001b. "Statistical Modeling: The Two Cultures." *Statistical Science* 16: 199–215.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. *Classification and Regression Tree.* New York: Chapman & Hall.

Bruininks, R. 1978. *Bruininks-Oseretsky Test of Motor Proficiency: Owner's Manual.* Circle Pines, MN: American Guidance Service.

Buhlmann, P., and B. Yu. 2002. "Analyzing Bagging." *Annals of Statistics* 30: 927–61.

Bylander, T. 2002. "Estimating Generalization Error on Two Class Datasets Using Out of Bag Estimates." *Machine Learning* 48: 287–97.

Cairney, J., J. A. Hay, B. E. Faught, and R. Hawes. 2005a. "Developmental Coordination Disorder and Overweight and Obesity in Children Aged 9–14 y." *International Journal of Obesity* 29 (4): 369–72.

Cairney, J., J. A. Hay, B. E. Faught, T. J. Wade, L. Corna, and A. Flouris. 2005b. "Developmental Coordination Disorder, Generalized Self-Efficacy Toward Physical Activity, and Participation in Organized and Free Play Activities." *Journal of Pediatrics* 147 (4): 515–20.

Cole, S. R., and M. A. Hernan. 2008. "Constructing Inverse Probability Weights for Marginal Structural Models." *American Journal of Epidemiology* 168 (6): 656–64.

D'Agostino Jr, R. B. 1998. "Propensity Score Methods for Bias Reduction in the Comparison of a Treatment to a Non-Randomized Control Group." *Statistics in Medicine* 17 (19): 2265–81.

Dietterich, T. G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning* 40: 1390157.

Drake, C. 1993. "Effects of Misspecification of the Propensity Score on Estimators of Treatment Effect." *Biometrics* 49: 1231–6.

Edwards, J., M. Berube, K. Erlandson, S. Haug, H. Johnstone, M. Meagher, S. Sarkodee-Adoo, and J. G. Zwicker. 2011. "Developmental Coordination Disorder in School-Aged Children Born Very Preterm and/or at Very Low Birth Weight: A Systematic Review." *Journal of Developmental and Behavioral Pediatrics* 32 (9): 678–87.

Flury, B. K., and H. Riedwyl. 1986. "Standard Distance in Univariate and Multivariate Analysis." *American Statistician* 40: 249–51.

Gentleman, R., and R. Ihaka. 2008. *R: A Language and Environment for Statistical Computing, R. D. C. Team.* Vienna, Austria: R Foundation for Statistical Computing.

Henderson, S., and D. Sugden. 1992. *Movement Assessment Battery for Children.* London: Psychological Corportation.

Lee, B. K., J. Lessler, and E. A. Stuart. 2010. "Improving Propensity Score Weighting Using Machine Learning." *Statistics in Medicine* 29 (3): 337–46.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by RandomForest." *R News* 2 (3): 18–22.

Meisels, S. J., D. B. Marsden, M. S. Wiske, and H. LW. 1997. *The Early Screening Inventory-Revised (ESI-R).* New York: Pearson Early Learning.

Missiuna, C., S. Moll, S. King, G. King, and M. Law. 2007. "A Trajectory of Troubles: Parents' Impressions of the Impact of Developmental Coordination Disorder." *Physical and Occupational Therapy in Pediatrics* 27 (1): 81–101.

Normand, S. T., M. B. Landrum, E. Guadagnoli, J. Z. Ayanian, T. J. Ryan, P. D. Cleary, and B. J. McNeil. 2001. "Validating Recommendations for Coronary Angiography Following Acute Myocardial Infarction in the Elderly: A Matched

Analysis Using Propensity Scores." *Journal of Clinical Epidemiology* 54 (4): 387–98.

Peters, A., and T. Hothorn. 2012. "Package 'Ipred'." *Improved Predictors*. CRAN.

Piek, J., N. Gasson, and J. Summers. 2008. "Motor Control and Coordination across the Lifespan." *Human Movement Science* 27 (5): 665–7.

Riethmuller, A. M., R. Jones, and A. D. Okely. 2009. "Efficacy of Interventions to Improve Motor Development in Young Children: A Systematic Review." *Pediatrics* 124 (4): e782–92.

Ripley, B. 2012. "Classification and Regression Trees" *Package 'Tree'*. CRAN.

Rosenbaum, P. R., and D. B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70: 41–55.

Rothman, K., S. Greenland, and T. Lash. 2008. "Casual Diagrams." In *Modern Epidemiology*, 3rd Edition, edited by S. Seigafuse, pp. 183–209. Philadelphia: Lippincott, Williams and Wilkins.

Rydz, D., M. I. Shevell, A. Majnemer, and M. Oskoui. 2005. "Developmental Screening." *Journal of Child Neurology* 20 (1): 4–21.

Setoguchi, S., S. Schneeweiss, M. A. Brookhart, R. J. Glynn, and E. F. Cook. 2008. "Evaluating Uses of Data Mining Techniques in Propensity Score Estimation: A Simulation Study." *Pharmacoepidemiology and Drug Safety* 17 (6): 546–55.

Strobl, C., J. Malloy, and G. Tutz. 2009. "An Introduction to Recursive Partitioning: Rationale, Application, and Characteristics of Classification and Regression Trees, Bagging and Random Forests." *Psychological Methods*, 14 (4): 323–48.

VanderWeele, T. J., and J. M. Robins. 2009. "Minimal Sufficient Causation and Directed Acyclic Graphs." *Annals of Statistics* 37 (3): 1437–65.

Watkins, S., M. Jonsson-Funk, M. A. Brookhart, S. A. Rosenberg, T. M. OShea, and J. Daniels. 2012. "Preschool Motor Coordination Following Physical and Occupational Therapy Services Among non-Disabled Very Low Birth Weight Children "Under Review.

Westreich, D., J. Lessler, and M. J. Funk. 2010. "Propensity Score Estimation: Neural Networks, Support Vector Machines, Decision Trees (CART), and Meta-Classifiers as Alternatives to Logistic Regression." *Journal of Clinical Epidemiology* 63 (8): 826–33.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.