Published in final edited form as:
J Public Health Dent. 2016 December ; 76(1): 17–29. doi:10.1111/jphd.12106.

Propensity Score Weighting: An Application to an Early Head Start Dental Study

Jacqueline M. Burgette, DMD[1,2,*], **John S. Preisser, PhD**[3], and **R. Gary Rozier, DDS, MPH**[2]

[1]Department of Pediatric Dentistry, School of Dentistry, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[2]Department of Health Policy and Management, UNC Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[3]Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

Abstract

Objectives—Non-randomized group assignment in intervention studies can lead to imbalances in pre-intervention covariates and biased effect estimates. We use propensity score estimation to account for such imbalances in an Early Head Start (EHS) dataset with rich pretreatment information. We compare propensity score results using standard logistic regression models (LRM) versus generalized boosted models (GBM).

Methods—We estimated propensity scores using 47 socio-demographic characteristics and EHS enrollment criteria obtained by parent interviews from a state-wide sample of 637 EHS and 930 Medicaid-matched control children. LRM and GBM were used to estimate propensity scores related to EHS enrollment. Performance of both approaches was evaluated via (1) measures of balance of pre-treatment covariate distributions between treated and control subjects; and (2) stability of propensity score weights measured by the effective sample size.

Results—Distributions of all variables were balanced for EHS and non-EHS groups using propensity score weights calculated with LRM and GBM. Compared to LRM, GBM resulted in better balance between treated and propensity score weighted control distributions. The effective sample size of the controls decreased from 930 subjects to 507 with GBM and to 335 with LRM.

Conclusion—Although propensity scores derived from GBM and LRM both effectively balanced observed pre-intervention covariates, GBM resulted in better covariate balance compared to LRM. GBM also resulted in a larger effective sample size of the control group compared to LRM. Propensity score weighting using GBM is an effective statistical method to reduce confounding due to imbalanced distributions of measured pre-intervention covariates in this EHS intervention study.

[*] Corresponding Author is Dr. Jacqueline M. Burgette, Department of Pediatric Dentistry, School of Dentistry, The University of North Carolina at Chapel Hill, 228 Brauer Hall, CB #7450, Chapel Hill, NC 27516. Tel.: 1-919-537-3955; Fax: 1-919-537-3950; jbur@email.unc.edu.

Keywords

Nonparametric statistics; Dental care for children; Health services research; Early Head Start

INTRODUCTION

Randomized trials are the gold-standard for evaluating interventions in clinical practice. However, random assignment of interventions often is not practical or ethical in the evaluation of public health or social programs. Experimental groups formed through non-random assignment can have an imbalance in the distribution of covariates, rendering the observed effects of any intervention the result of selection bias, not the intervention itself. Quasi- or non-experimental designs are the most common ones used in studies evaluating the effects of participation in early education programs (1,2). Quasi-experimental studies attempt to control selection bias by inclusion of a comparison group with characteristics similar to the intervention group. Martin and colleagues (3), for example, selected control children for a retrospective cohort study of the effects of Head Start on dental use from Medicaid files by matching on race, age, sex and urban county of residence. This design limits the investigators' ability to control for group imbalances in variables that could affect dental use because of the small number of covariates available for study.

A number of methods have been proposed to account for selection bias in non-randomized studies (4). One of these involves weighting the control observations to balance the distributions of pre-intervention characteristics (5, 6). Most commonly, the weight used in this context depends on the propensity score, which is defined as the probability of being in the intervention group based on a selection of observed characteristics (7). It is a "balancing score" such that – conditional on the propensity score – the distribution of pretreatment covariates is the same in the intervention and control groups. Therefore, controlling for the propensity score corrects observed imbalances in pretreatment characteristics among experimental groups (8, 9).

Propensity scores can be estimated using parametric models, such as logistic regression models (LRM), and nonparametric models, such as generalized boosted models (GBM). Some researchers prefer nonparametric models because they provide more flexibility in capturing complicated relationships between the pretreatment covariates and the treatment indicator (10). Estimating propensity scores using generalized boosted models has been well-developed in the statistical literature (7–9, 11–16), and has become popular in many fields including education, health services, psychology and drug addiction (4– 6, 10, 17–24). Arguably, propensity scores in general have been used in only a few studies in dentistry (25–32); and more modern approaches to propensity score estimation, such as generalized boosted models, have not been used at all.

In this study, we implement propensity score weighting to control for selection bias into an Early Head Start (EHS) program evaluation and compare results using two modeling approaches: LRM and GBM. EHS is a federally-funded, community-based early education and childcare program that offers a number of health services for pregnant women and children from birth to 3 years of age in low-income families (33, 34). The evaluation project for the North Carolina EHS program, known as the ZOE (Zero-Out Early Childhood Caries), is designed as a non-randomized, pretest-posttest nested cohort-control group cluster trial and aims to determine the effects of EHS on oral health outcomes in enrolled children.

We recruited 25 EHS programs with centers in 41 counties of the state and enrolled child-parent dyads during two sequential school years beginning in 2010. Only one child was selected from each family according to specified criteria, which included the youngest child in families with more than one. A control group was selected from Medicaid children matched to the EHS group on age, residential ZIP code and whether English or Spanish was the preferred language. ZOE has the potential for two primary sources of selection bias. Compared to control subjects, those families that choose to apply for EHS enrollment might have characteristics that predispose them to behaviors that differentially affect oral health outcomes. Federal guidelines require that certain criteria, such as income levels, be met for families to be eligible for the program. Within the broad guidelines, EHS programs also must try to meet local community needs, which means that programs can prioritize enrollment criteria differently and enroll children with characteristics that would affect oral health outcomes. In this study, many of the same factors that lead families to apply for enrollment or are used by EHS programs to enroll families into EHS might also influence oral health, and therefore are potential confounders. In this study we compare the performance of propensity score weights generated by LRM versus GBM in controlling for selection bias that might affect ZOE study results.

METHODS

Data Sources and Variable Selection

This study uses data collected as part of the ZOE study, an evaluation of the impact of preventive dental services provided by EHS programs and primary care physicians on the oral health status of children enrolled in North Carolina EHS programs. Parents of children in EHS (n=636) and community controls (n=931) were interviewed in the child’s first year of life and approximately 24-month later at EHS program end. In-person, interviews were conducted in English or Spanish with an adult family member by interviewers trained in structured interviewing techniques. The interviews lasted about 60 minutes on average and included a range of topics that could be affected by EHS, including oral health-related knowledge about child care, oral health behaviors, dental visits, and oral health-related quality of life. The interview also included sociodemographic characteristics and a number of items designed to obtain individual-level information on federal and program enrollment criteria for EHS. These items were based on questions included in family application forms used by all EHS programs in the state at the time of the study.

We chose covariates for the propensity score analysis that we hypothesized would be correlated with the treatment assignment, EHS, as well as the dental outcomes (5, 10). From the rich selection of covariates measured prior to the EHS intervention we chose 25 socio-demographic variables that can predispose families to enroll in EHS (Appendix A), and 22 EHS selection criteria that EHS programs use to determine enrollment eligibility (Appendix B).

Propensity Score Method

GBM is an automated algorithm that uses regression trees and boosting (10, 35). A regression tree is a “forward stagewise additive algorithm” in which a split, or tree branch, is added with each model iteration to achieve the greatest increase of the likelihood among all possible splits (36, 37). The recursive splits automatically accommodate non-linear and interactive effects (10, 37). While each tree partition is simple in that each split depends on a single covariate, the cumulative effect of many splits can describe complex relationships between the covariates and group assignment (37). The modern statistical technique of additively combining many simple models into a more complex model is called boosting, and can result in better predictive performance than any of the simple models can provide (11). With boosting, GBM retains the attractive features of regression trees, and can successfully incorporate a large number of measured covariates. When multiple trees are combined, it results in a smoother fit and better prediction than simple regression trees alone and resists the tendency of tree-based models to over-fit the data (11). GBM also stabilizes propensity score weighted estimators by flattening out at the extreme values including those close to zero or one (5, 10). Each of these GBM characteristics is desirable in the context of propensity score estimation.

Average effect of Treatment on the Treated—We considered the average effect of treatment on the treated (ATT), which is “the average effect that would be seen if everyone in the treated group received the treatment, compared to if no one in the treated group received the treatment” (5). Our propensity score analysis employed ATT over Average Treatment Effect, which describes the mean difference in outcomes if all individuals in the population had received the treatment versus if all individuals had received the control, for two reasons (17). First, the treatment, EHS, is not intended for the entire population of NC children, but select, low-income children who would most benefit from EHS services. Second, we have no information on how high-income families behave in EHS because they are not eligible for enrollment. In our application of propensity scores in this study, the treated group (EHS) children received a weight equal to one and the control (non-EHS) children received a weight described by the relationship (38):

$$\text{Propensity score weight}=\text{Propensity score}/(1-\text{Propensity score}). \quad (1)$$

This approach is defined as weighting by the odds. These weights adjust the control group so that its pretreatment covariate distributions match those of the treatment group. Subtracting the weighted mean outcome observed among the controls from the mean outcome among the treated estimates the ATT (5).

Weights, such as those applied to our control sample, can reduce the precision of our statistical estimates so that our analyses behave as though we had a smaller sample of control subjects. The “effective sample size” (ESS) quantifies this reduction of the control group and is given by

$$ESS=(\Sigma w_i)^2/\Sigma w_i^2 \quad (2)$$

where w_i is the propensity score weight for the *i*th child (9). An analysis with a control ESS of a given size behaves similarly in terms of statistical power as an unweighted analysis with that number of control observations.

Balance Measures—We used two summary statistics to measure balance between the EHS and non-EHS groups for each pretreatment covariate: the absolute standardized mean difference (ASMD) and the Kolmogorov-Smirnov distance (KS) (10). The ASMD equals the absolute value of the difference between the weighted mean for treatment group and the weighted mean for the control group divided by the unweighted standard deviation of the treatment group. For ATT analyses, the KS is the maximum vertical distance between the unweighted empirical cumulative distribution function (EDF) for the treatment sample and weighted EDF for the control sample.

Propensity Score Modeling—We estimated propensity scores using LRM and GBM and applied them in the form of a propensity score weight (Equation 1). We created the LRM with no interactions among covariates or functional forms higher than linear terms. We avoided dropping observations in the LRM by performing mean imputation. To handle item nonresponse and data missing by design in the GBM, we employed a missing indicator approach by balancing on both the observed values of each covariate as well as missing data patterns for each covariate, which is the default of the software described by Ridgeway and colleagues (2013) (35). Although the overall percentage of missing data is low at 6.9%, they were concentrated in three covariates in excess of 50% missing by design as a result of a skip pattern in which no response from the interviewee was intended. No cases had complete data, however we kept all observations because the EHS treatment status was observed for every child-parent dyad.

We used the maximum ASMD as the stopping rule for the complexity of the GBM. We then examined both the mean and the maximum of the ASMD and KS across covariates as a metric of propensity score performance for both the LRM and GBM. We considered ASMD values greater than 0.2 to indicate moderate imbalance (39) and a KS greater than 0.10 as an indication of imbalance (35). All modeling was performed in the R software environment. We fit the GBM using the Toolkit for Weighting and Analysis of Nonequivalent Groups (TWANG) package (10, 35).

Propensity score weighting by the odds using GBM was implemented in this study for the following reasons: 1. to avoid dropping observations since GBM is able to estimate propensity scores even for cases with missing data elements; 2. evidence that GBM estimation of propensity scores works particularly well when applied to weighting of the odds (5); and 3. the criterion (ASMD stopping rule) used to select the GBM complexity in the TWANG package assumed that the propensity scores will be used for weighting (35).

Evaluation of Propensity Score Performance—We used three criteria to evaluate the success of our propensity score weighting of ZOE data. In order of importance, these criteria were: (1) covariate balance between treatment and control groups, (2) relatively homogenous propensity score weights, and (3) effective sample size of the control group.

RESULTS

Covariate Balance

We detected a substantial difference for 12 pre-treatment variable means between EHS and non-EHS groups, indicated by an imbalance in pre-treatment ASMD greater than 0.2 (Table 1, Figure 1 unweighted analysis). Of the eight socio-demographic covariates and four EHS selection criteria covariates that were imbalanced, the observed pre-treatment characteristic with the greatest difference was non-Hispanic White child race and ethnicity, with the non-EHS control group having substantially more non-Hispanic White children (Appendix A). After weights derived from the propensity scores were applied, all pre-treatment differences between the EHS and non-EHS groups diminished to a substantially improved ASMD below 0.2 (Figure 1 weighted analysis). Therefore, variables that were imbalanced in the unweighted analysis became balanced in the weighted propensity score analysis for both the LRM (Table 2) and GBM models (Figure 1 and Table 2). Distributions of all covariates were balanced successfully using propensity score weights calculated with both LRM and GBM (Table 2). For example, 18.5% of children were non-Hispanic White in the EHS group, which was significantly different from 35.6% of children in the non-EHS group ($p<0.05$). Following the application of GBM propensity score weights, the percentage of children with non-Hispanic White race and ethnicity drops to 20.0% in the non-EHS group, which is not significantly different from the 18.5% in the EHS group ($p\geq0.05$) (Appendix A).

Although both GBM and LRM propensity score weights resulted in balance in the pre-treatment variables between the EHS and non-EHS groups as measured by the 0.2 ASMD cutoff, GBM resulted in better balance than LRM. Compared to LRM, GBM performed substantially better with preferred smaller values for maximum ASMD, a statistical measure of balance (Table 2). The maximum ASMD was 0.194 for the LRM compared to 0.129 for the GBM (Table 2). GBM and logistic regression performed similarly for the other three statistical measures of balance (maximum KS, mean ASMD and mean KS) (Table 2).

Propensity Score Estimation and Propensity Score Weighting by the Odds

Substantial overlap was observed in the overall distribution of estimated propensity scores between the EHS and non-EHS groups for both GBM and LRM analyses (Figure 2). However, the scores from the LRM were more dispersed than the GBM estimates, with more homogeneity in propensity scores being preferred (so long as balance is achieved). For the non-EHS group, propensity scores ranged from 2.22e–16 to 0.92 for LRM compared to 0.044 to 0.85 for GBM (Figure 2). Compared to LRM, GBM propensity scores for the non-EHS group were not as close to zero or one (Figure 2). For ATT analyses, propensity score weights depend only on the non-EHS propensity scores; however, the propensity scores for the EHS group were also less variable when using GBM.

The LRM propensity score estimates that nearly equal one result in very large weights. The range of the propensity score weights in the GBM analysis ranged from 0.046 to 5.87 compared to 2.22e–16 to 11.00 in the LRM analysis. Propensity scores with less variability, as seen in GBM compared to LRM, resulted in more homogenous propensity score weights (Equation 1) and a larger effective sample size of the control group (Equation 2).

Effective Sample Size of the Control Group

The effective sample size for the GBM weights is 507, down from the nominal size of 930. This finding means that the propensity score weighted analyses have approximately the same statistical power as an unweighted sample with 507 non-EHS children. The effective sample size for LRM is lower, at 335 (Table 2).

DISCUSSION

Even though the EHS and non-EHS groups differed considerably at baseline on characteristics that might bias evaluation results, propensity score weighting balanced the groups on all 47 pretreatment variables. Covariate imbalance between treatment and control groups is a common problem in non-randomized research designs where researchers seek to identify the causal effects from observational studies. The positive findings of this study are important because lack of randomization is a wide-spread challenge in dental research, particularly the evaluation of public health practice.

Compared to LRM, propensity score weights derived from GBM resulted in better balance, indicating that GBM is more capable of removing bias in baseline differences. While both LRM and GBM achieve covariate balance with ASMD values below the 0.2 cutoff, the maximum ASMD for LRM (0.194) was marginally below 0.2 while the maximum ASMD for GBM (0.129) was substantially below 0.2, signifying better covariate balance with GBM.

GBM also resulted in a larger effective sample size of the control group, which results in better statistical power in the weighted analyses to detect differences between EHS and non-EHS groups in the ZOE study. All of these findings suggest that propensity score analysis using weights derived from GBM propensity score estimation is a promising technique for analysis of the impact of EHS on oral health outcomes.

To our knowledge, this is the first study to use GBM propensity score weighting in dental research. However, it is increasingly used in other disciplines (12, 18, 19). Several studies have shown improved results with propensity scores derived from GBM compared to LRM, including smaller standard errors for the treatment effects and better covariate balance (5, 10, 13, 14).

While propensity score analyses can control for selection into the EHS group using observable characteristics, they cannot be used to control for selection based on characteristics that are unobserved (20, 21). We anticipated the potential for imbalances in EHS groups even after matching control subjects from other low-income families by including a large number of covariates and enrollment criteria in the baseline interview. Nevertheless, the unmeasured confounders can persist and bias treatment effect estimates, violating the assumption of “strong ignorable treatment assignment” (7). A limitation of both the LRM and GBM models, and propensity score analyses in general, is that hidden bias may remain due to unmeasured confounders. Robustness to a large degree of hidden bias due to unmeasured confounders can be determined in the final EHS outcome model by performing a sensitivity analysis (10, 15, 22, 23).

An alternative to propensity score analysis is instrumental variable analysis. However, it would be challenging to identify a valid instrument in ZOE because many covariates were included in baseline parent interviews for their pertinence to oral health outcomes. While the inclusion of covariates correlated with the outcome is undesirable as an instrumental variable, it is preferred in propensity score analysis (16).

A limitation of this study is that we did not include higher-ordered functional forms and interaction terms in the LRM, which could improve balance between the EHS and non-EHS groups. However, the LRM already achieved acceptable covariate balance; and a more complex LRM would likely shrink the effective sample size of the control group beyond what we observed in our analysis. We also designed the LRM to behave favorably with respect to the study criteria by using mean imputation for missing data to diminish differences between the EHS and non-EHS groups while increasing the available sample size.

Although GBM and LRM both effectively balanced pre-intervention covariates, GBM resulted in better balance compared to LRM as determined by: (1) better values on statistical balance measures, (2) greater homogeneity in propensity scores weights, and (3) a larger effective sample size for the control group. Overall, we conclude that if there is a rich set of pre-treatment variables that are correlated with both the treatment and outcome, then GBM may provide better balance and effective sample size than LRM propensity scores.

REFERENCES

1. Barnett, WS. Preschool Education Studies: A Bibliography Organized by Research Strengths. New Brunswick: National Institute for Early Education Research; 2007. Available from: http://nieer.org/publications/nieer-working-papers/preschool-education-studies-bibliography-organized-research [cited 2015 Feb 20]
2. National Forum on Early Childhood Program Evaluation. Early Childhood Program Evaluations: A Decision-Maker's Guide. Cambridge: National Forum on Early Childhood Program Evaluation; 2007. Available from: http://www.developingchild.harvard.edu [cited 2015 Feb 20]
3. Martin AB, Hardin JW, Veschusio C, Kirby HA. Differences in dental service utilization by rural children with and without participation in Head Start. Pediatr Dent. 2012 Sep-Oct;34(5):107–111. [PubMed: 23211894]
4. Pearl, J. Causality: Models, Reasoning, and Inference. 2nd edition. New York: Cambridge University Press; 2009.
5. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. Psychol Methods. 2010 Sep; 15(3):234–249. [PubMed: 20822250]
6. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis. 2007; 15:199–236.
7. Rosenbaum PR, Rubin DB. The Central Role of the Propensity Score in Observational Studies for Causal Effects. Biometrika. 1983; 70:41–55.
8. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician. 1985; 39:33–38.

9. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med. 1998 Oct 15; 17(19):2265–2281. [PubMed: 9802183]

10. McCaffrey DF, Ridgeway G, Morral A. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. Psychological Methods. 2004 Dec; 9(4):403–425. [PubMed: 15598095]

11. Friedman JH, Hastie T, Tibshirani R. Additive logistic regression: A statistical view of boosting. Annals of Statistics. 2000; 28:337–374.

12. Karwa V, Slavković AB, Donnell ET. Causal inference in transportation safety studies: Comparison of potential outcomes and causal diagrams. The Annals of Applied Statistics. 2011; 5:1428–1455.

13. Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med. 2010 Feb 10; 29(3):337–346. [PubMed: 19960510]

14. Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. PLoS ONE. 2011 Mar 31.6(3):e18174. [PubMed: 21483818]

15. Rosenbaum PR, Rubin DB. Assessing Sensitivity to an Unobserved Binary Covariate in an Observational Study with Binary Outcome. Journal of the Royal Statistical Society. Series B (Methodological). 1983; 45:212–218.

16. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. Biometrics. 1993; 49:1231–1236.

17. Hirano K, Imbens G. Estimation of Causal Effects using Propensity Score Weighting: An Application to Data on Right Heart Catheterization. Health Services and Outcomes Research Methodology. 2001; 2:259–278.

18. Pollack CE, Griffin BA, Lynch J. Housing affordability and health among homeowners and renters. Am J Prev Med. 2010 Dec; 39(6):515–521. [PubMed: 21084071]

19. Cohen DA, Golinelli D, Williamson S, Sehgal A, Marsh T, McKenzie TL. Effects of park improvements on park use and physical activity: policy and programming implications. Am J Prev Med. 2009 Dec; 37(6):475–480. [PubMed: 19944911]

20. Brooks JM, Ohsfeldt RL. Squeezing the balloon: propensity scores and unmeasured covariate balance. Health Serv Res. 2013 Aug; 48(4):1487–1507. [PubMed: 23216471]

21. Toh S, Gagne JJ, Rassen JA, Fireman BH, Kulldorff M, Brown JS. Confounding adjustment in comparative effectiveness research conducted within distributed research networks. Med Care. 2013 Aug; 51 Suppl 3(8):S4–S10. [PubMed: 23752258]

22. Haviland A, Nagin DS, Rosenbaum PR. Combining propensity score matching and group-based trajectory analysis in an observational study. Psychol Methods. 2007 Sep; 12(3):247–267. [PubMed: 17784793]

23. Schafer JL, Kang J. Average causal effects from nonrandomized studies: a practical guide and simulated example. Psychol Methods. 2008 Dec; 13(4):279–313. [PubMed: 19071996]

24. Lanza ST, Moore JE, Butera NM. Drawing causal inferences using propensity scores: a practical guide for community psychologists. Am J Community Psychol. 2013 Dec; 52(3–4):380–392. [PubMed: 24185755]

25. Murphy DA, Harrell L, Fintzy R, Belin TR, Gutierrez A, Vitero SJ, Shetty V. A Comparison of Methamphetamine Users to a Matched NHANES Cohort: Propensity Score Analyses for Oral Health Care and Dental Service Need. J Behav Health Serv Res. 2014 Nov 15. [Epub ahead of print].

26. Mac Giolla Phadraig C, McCallion P, Cleary E, McGlinchey E, Burke E, McCarron M, Nunn J. Total tooth loss and complete denture use in older adults with intellectual disabilities in Ireland. J Public Health Dent. 2014 Oct 15. [Epub ahead of print].

27. Kranz AM, Rozier RG, Preisser JS, Stearns SC, Weinberger M, Lee JY. Preventive Services by Medical and Dental Providers and Treatment Outcomes. J Dent Res. 2014 Jun 2; 93(7):633–638. [PubMed: 24891593]

28. Brickhouse TH, Haldiman RR, Evani B. The impact of a home visiting program on children's utilization of dental services. Pediatrics. 2013 Nov. Suppl 2(132):S147–S152. [PubMed: 24187117]

29. Sen S, Sumner R, Hardin J, Barros S, Moss K, Beck J, Offenbacher S. Periodontal disease and recurrent vascular events in stroke/transient ischemic attack patients. J Stroke Cerebrovasc Dis. 2013 Nov; 22(8):1420–1427. [PubMed: 23910516]

30. Chaudhari M, Hubbard R, Reid RJ, Inge R, Newton KM, Spangler L, Barlow WE. Evaluating components of dental care utilization among adults with diabetes and matched controls via hurdle models. BMC Oral Health. 2012 Jul 9.12:20. [PubMed: 22776352]

31. Stearns SC, Rozier RG, Kranz AM, Pahel BT, Quiñonez RB. Cost-effectiveness of preventive oral health care in medical offices for young Medicaid enrollees. Arch Pediatr Adolesc Med. 2012 Oct; 166(10):945–951. [PubMed: 22926203]

32. Shetty V, Mooney LJ, Zigler CM, Belin TR, Murphy D, Rawson R. The relationship between methamphetamine use and increased dental disease. J Am Dent Assoc. 2010 Mar; 141(3):307–318. [PubMed: 20194387]

33. Early Head Start National Resource Center. Washington, DC: Office of Head Start, Administration for Children and Families, U.S. Department of Health and Human Services; 2014. About Early Head Start [Internet]. Available from: https://eclkc.ohs.acf.hhs.gov/hslc/tta-system/ehsnrc/about-ehs [cited 2015 Feb 20]

34. Washington, DC: Head Start Bureau, Administration on Children, Youth and Families, U.S. Department of Health and Human Services; Head Start Program Performance Standards and Other Regulations [Internet]. Available from: http://eclkc.ohs.acf.hhs.gov/hslc/standards] [cited 2015 Feb 20]

35. Ridgeway G, McCaffrey D, Morral A, Griffin BA, Burgette L. TWANG: Toolkit for Weighting and Analysis of Nonequivalent Groups [Internet]. R package version 1.3-21. 2013 Available from: http://CRAN.R-project.org/package=twang.

36. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. 2nd ed.. New York: Springer Science and Media Inc; 2009.

37. Breiman, L.; Friedman, JH.; Olshen, RA.; Stone, CJ. Classification and regression trees. Belmont: CRC Press; 1984.

38. Hirano K, Imbens G, Ridder G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. Econometrica. 2003; 71:1161–1189.

39. Cohen, J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed.. Hillsdale: Lawrence Erlbaum; 1998.

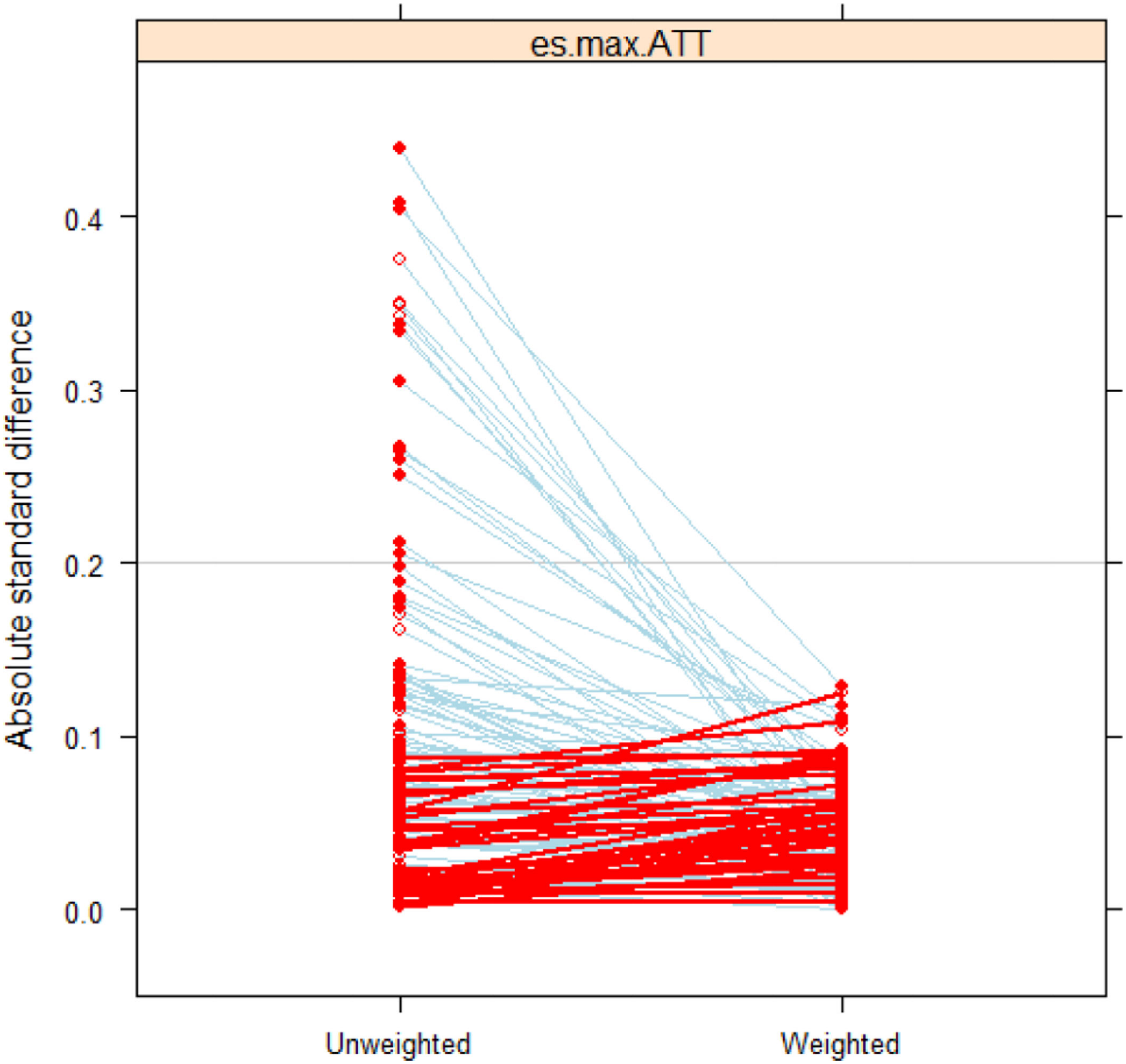

Figure 1. Absolute standard mean difference in pre-treatment characteristics before (unweighted) and after (weighted) Generalized Boosted Model propensity score weights

An open circle indicates an insignificant difference in the mean of a particular pre-treatment variable between the treated and control groups (p≥0.05). A closed circle indicates a significant difference in the mean of a particular pre-treatment variable between the treated and control groups (p<0.05). Significance is defined via a two-sample t-test for continuous variables and a chi-square test for categorical variables at a significance level of 0.05. The left side of the figure depicts covariate balance before applying propensity score weights. The right side of the figure depicts covariate balance after applying propensity score weights. Specific mean values for each variable, as well as an indication of variables with an absolute standard mean difference greater than 0.2, are available in Appendix A and B.

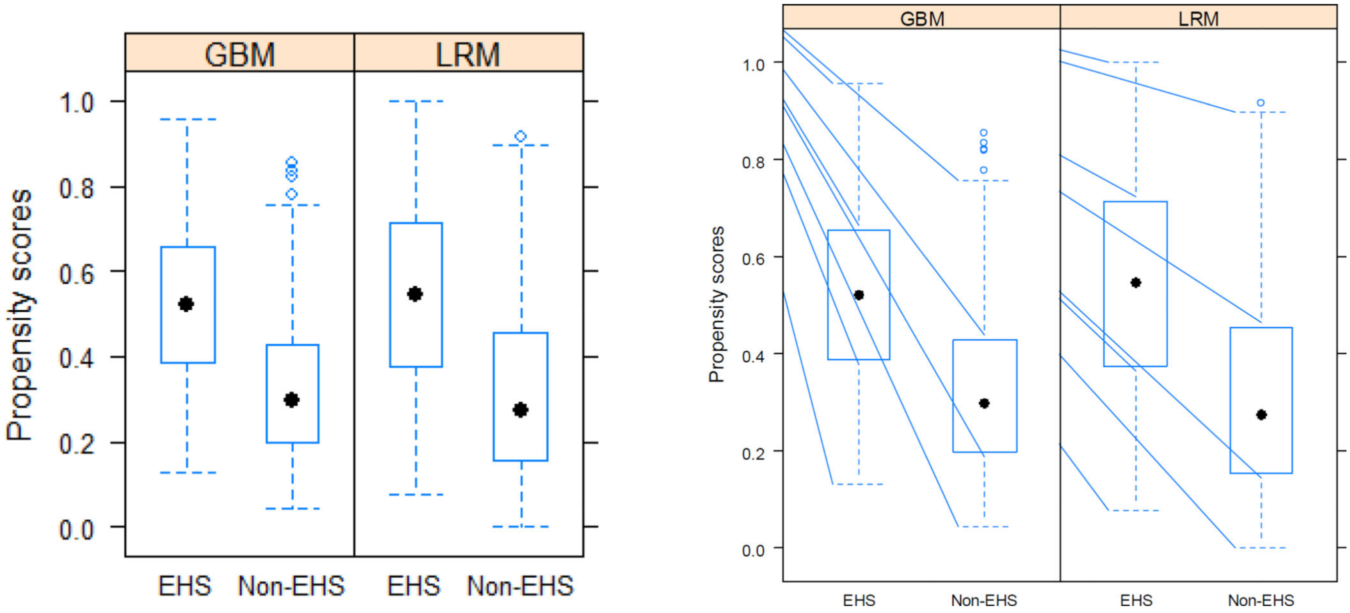

Figure 2. Boxplots of Propensity Score Distribution for Early Head Start and non-Early Head Start using Generalized Boosted Model (GBM) and Logistic Regression Model (LRM)

The boxes mark the first (25th percentile) and third (75th percentile) quartiles of the propensity scores with solid circle at the median. The dashed lines extending from the boxes indicate the medians plus and minus 1.5 times the interquartile range. Data points beyond 1.5 times the interquartile range are represented by an open circle.

Table 1

Pre-treatment variables that were imbalanced between the EHS and non-EHS groups prior to propensity score analysis

Socio-demographic Variable	*On average, EHS Group had:*
Brothers or sisters	More brothers and sisters
Number in household under 5 years-old	More people under 5 years-old in the household
Caregiver marital status	More single/never married caregivers
Caregiver Race and ethnicity: Non-Hispanic White.	Fewer non-Hispanic White caregivers
Caregiver Race and ethnicity: Non-Hispanic African-American	More African-American caregivers
Caregiver Race and ethnicity: Non-Hispanic single other race	Fewer non-Hispanic single other race caregivers
Child Race and ethnicity: Non-Hispanic White	Fewer non-Hispanic White children
Child Race and ethnicity: Non-Hispanic African American	More African-American children
Early Head Start Selection Criteria	*On average, EHS Group had:*
Does any of your household receive Food Stamps?	Received more Food Stamps
Does any of your household receive Child care subsidy or education assistance?	Received more child care subsidy or education assistance
Does any member of your household receive Housing assistance?	Received more housing assistance
Caregiver in school or training	More caregivers were in school or training

Table 2

Balance Measures and Sample Sizes for unweighted and propensity score weighted samples using logistic regression and generalized boosted models

	EHS Sample Size	Non-EHS Sample Size	EHS Effective Sample Size	Non-EHS Effective Sample Size	Maximum ASMD†	Mean ASMD†	Maximum K-S‡distance	Mean K-S‡ distance
Logistic Regression Model								
Unweighted Model	637	930	637	930	0.439	0.099	0.230	0.027
Propensity Score Model	637	930	637	335	0.194	0.035	0.083	0.008
Generalized Boosted Model								
Unweighted Model	637	930	637	930	0.439	0.097	0.229	0.025
Propensity Score Model	637	930	637	507	0.129	0.050	0.071	0.010

†ASMD = absolute standardized mean difference, which is the difference in mean for treatment and control for each covariate divided by the standard deviation of the treated group. Maximum ASMD is the largest difference in mean for treatment and control for each covariate divided by the standard deviation of the treated group. Mean ASMD is the average difference in mean for treatment and control for each covariate divided by the standard deviation of the treated group.

‡K-S = Kolmogorov-Smirnov distance is the difference in the empirical cumulative distribution functions between the treatment and control for each covariate. Maximum Kolmogorov-Smirnov distance is the largest difference in the cumulative distribution functions between the treatment and control groups for each covariate. Mean Kolmogorov-Smirnov distance is the average difference in the cumulative distribution functions between each covariate.

Note: Differences in the ASMD and K-S for the unweighted LRM and GBM are due to the imputation of missing variables in the LRM.