I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17080

# Breastfeeding and Child Development Outcomes across Early Childhood and Adolescence: Doubly Robust Estimation with Machine Learning

Md Mohsan Khudri
Andrew Hussey

JUNE 2024

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17080

# Breastfeeding and Child Development Outcomes across Early Childhood and Adolescence: Doubly Robust Estimation with Machine Learning

**Md Mohsan Khudri**
*Austin College*

**Andrew Hussey**
*University of Memphis and IZA*

JUNE 2024



ISSN: 2365-9793

**IZA – Institute of Labor Economics**

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

# ABSTRACT

## Breastfeeding and Child Development Outcomes across Early Childhood and Adolescence: Doubly Robust Estimation with Machine Learning*

Using data from the Panel Study of Income Dynamics, we estimate the impact of breastfeeding initiation and duration on multiple cognitive, health, and behavioral outcomes spanning early childhood through adolescence. To mitigate the potential bias from misspecification, we employ a doubly robust (DR) estimation method, addressing misspecification in either the treatment or outcome models while adjusting for selection effects. Our novel approach is to use and evaluate a battery of supervised machine learning (ML) algorithms to improve propensity score (PS) estimates. We demonstrate that the gradient boosting machine (GBM) algorithm removes bias more effectively and minimizes other prediction errors compared to logit and probit models as well as alternative ML algorithms. Across all outcomes, our DR-GBM estimation generally yields lower estimates than OLS, DR, and PS matching using standard and alternative ML algorithms and even sibling fixed effects estimates. We find that having been breastfed is significantly linked to multiple improved early cognitive outcomes, though the impact reduces somewhat with age. In contrast, we find mixed evidence regarding the impact of breastfeeding on non-cognitive (health and behavioral) outcomes, with effects being most pronounced in adolescence. Our results also suggest relatively higher cognitive benefits for children of minority mothers and children of mothers with at least some post-high school education, and minimal marginal benefits of breastfeeding duration beyond 12 months for cognitive outcomes and 6 months for non-cognitive outcomes.

**JEL Classification:** I12, I18, J13, J24, C21, C63

**Keywords:** breastfeeding, human capital, cognitive and non-cognitive outcomes, doubly robust estimation, machine learning

**Corresponding author:**
Andrew Hussey
Department of Economics
Fogelman College of Business and Economics
University of Memphis
3720 Alumni Ave
Memphis, TN 38152
USA

E-mail: ajhussey@memphis.edu

* We appreciate the insightful comments from Jamin Speer, Han Yu, David Kemme and participants at the 2023 Southern Economic Association Meetings. Any error is ours.

## 1| Introduction

A large body of research underscores the importance of health capital investment in infants and toddlers on their developmental and socioeconomic outcomes in early and subsequent stages of life (Cunha, Heckman, & Schennach, 2010; Francesconi & Heckman, 2016; Heckman & Mosso, 2014; Maluccio et al., 2009; Walters, 2015). Breastfeeding, in particular, has been extensively documented to confer numerous benefits for children.[1] Several studies have pointed to improved health outcomes, including reduced risks of asthma, diabetes, obesity, ear infections, disability, and stunting (Adair & Guilkey, 1997; Brennan et al., 2004; Dyson et al., 2006; Haines & Kintner, 2008; Ip et al., 2007; Wehby, 2014; Victora et al., 2016).[2] More consistently, strong associations between breastfeeding and child cognitive outcomes have been found (Borra et al., 2012; Belfield & Kelly, 2012; Del Bono -& Rabe, 2012; Evenhouse &Reilly, 2005; Kramer et al., 2001; Fitzsimons & Vera-Hernández, 2022), along with higher educational attainment (Mohammed et al., 2023; Rees & Sabia, 2009; Victora et al., 2015). Given these findings and persisting gaps in breastfeeding by family income, education, and race (Diaz et al., 2023), policies targeting breastfeeding have the potential to significantly impact social and economic inequality in the intergenerational transmission of human capital (Almond et al., 2018).

Much of the evidence of the impacts of breastfeeding has necessarily come from observational studies. Since the act of breastfeeding is a personal choice and influenced by one's specific circumstances, deriving causal estimates from observational data requires both rich data

[1] While not reviewed here, there is a large literature relating to the mechanisms of why breastfeeding may lead to improved developmental outcomes. One of two main channels is the nutritional content of breast milk versus infant formulas, including particular fatty acids, as well as transference of antibodies for immune protection (Martin et al., 2016). The other main channel is through increased skin-to-skin contact between mothers and babies, which may stimulate oxytocin release and improve mother-child bonding (Bigelow & Power, 2020).

[2] Other studies observe no significant impacts of breastfeeding on health outcomes in childhood (Del Bono & Rabe, 2012; Evenhouse & Reilly, 2005; Fitzsimons & Vera-Hernandez (2022).

and correctly specified empirical models in order to control for this selection.[3] This paper uses data from the Panel Study of Income Dynamics (PSID) to estimate the relationships between breastfeeding and several cognitive, health, and behavioral outcomes of children in the United States at two points in time. Along with propensity score matching (PSM) estimators, we use a doubly robust (DR) estimation procedure involving propensity scores (PS), which allows for unbiased estimates even when either the selection equation or outcome model is misspecified. For both methods, our focus is on accurate estimation of the PS (i.e., the likelihood of being breastfed). We use a battery of supervised machine learning algorithms (ML) to estimate the PS, demonstrably improving estimation upon traditional approaches.

Despite the infeasibility of directly experimentally manipulating breastfeeding activity, a handful of studies in the breastfeeding literature use experimental or natural experimental designs. Mostly notably, Kramer et al. (2008a, b) utilize a randomized trial of a breastfeeding promotion intervention in Belarus. In this "intention to treat" framework, they find that exclusive breastfeeding improves cognitive and health outcomes of children. A recent study by Fitzsimons & Vera-Hernández (2022) cleverly exploits a natural experiment involving differences in hospital staffing on weekends versus weekdays in the United Kingdom, also finding large positive effects of breastfeeding on cognitive outcomes of children – especially those of lower educated women.[4] Despite its obvious strengths, there are some downsides to this natural experimental approach. First, the settings in which such an approach is possible are quite limited, making generalizations

[3] Selection into breastfeeding is generally assumed (or observed) to be positive. Supporting this, Raissian & Su (2018) find that prenatal breastfeeding *intentions* are positively related to infant health outcomes, regardless of whether the mother ever initiated breastfeeding. Part of this association appears to be due to differences in health knowledge (prior to giving birth) between women intending versus not intending to breastfeed.

[4] Other studies use variation or changes in policies in an attempt to estimate causal impacts. Baker & Milligan (2008) exploit changes in Canadian maternal leave mandates, which they find to positively impact breastfeeding duration. Del Bono & Rabe (2012) use variation in breastfeeding support policies across UK hospitals, finding large positive impacts on child cognitive development.

to other countries or demographics difficult. Second, this approach results in the estimation of a Local Average Treatment Effect (LATE). For example, according to Fitzsimons & Vera-Hernández (2022, p. 360): "…Our estimates are only applicable to compliers, who are relatively well-off mothers (amongst those with relatively low education) and those who experienced some complications during delivery, and we cannot extrapolate our results to other groups of the population."

In the absence of an experimental or natural experimental setting, researchers tend to use one of three approaches: (1) Regression (OLS, logit/probit, etc.) including few or many controls; (2) sibling fixed effects, which helps to eliminate unobserved family environment and some genetic characteristics from confounding the estimated treatment effect;[5] and (3) PSM.[6] The latter approach is a particularly common method used to reduce potential confounding biases in observational studies estimating average treatment effects. A semi-parametric approach, PSM attempts to mimic an experimental setting by comparing outcomes among treated and untreated individuals, where the two groups are made observationally similar in terms of their PS. The ability of PSM to generate unbiased estimates rests on accurate estimation of the PS, the probability of treatment (Rosenbaum & Rubin, 1983). A typical approach is for researchers to select variables to include in a large, flexible-form logit or probit, the specification of which is selected to balance the PS of treated and control units. To the extent that the PS formulation is misspecified, the resulting PSM estimates will be biased.

[5] Evenhouse & Reilly (2005), Rees & Sabia (2009), and Wehby (2014) incorporate sibling fixed effects in their analyses. While offering a clear improvement, a disadvantage of this approach is that it relies solely on multi-child households that, for whatever reason, breastfed at least one child but did not breastfeed at least one other.
[6] For example, in the literature on impacts of breastfeeding, PSM is used by Belfield & Kelly (2012), Borra, Iacovou, & Sevilla (2012), McCrory & Layte (2011), and Rothstein (2013).

Methodologically, this study utilizes PS in PSM, as well as DR estimation, which allows for separating the outcome specification from the selection (PS) specification, resulting in unbiased treatment effects estimates if either (but not both) specification is incorrect. A novel aspect of our paper is that we use ML methods to improve estimation of the PS. Owing to the significant heterogeneity of the underlying data, ML techniques are particularly effective in predicting human behavior and health outcomes (Kleinberg et al., 2018).[7] The benefits of ML techniques for regression have recently become more pronounced in the economics literature, indicating the problems in economics where ML methods are robust and emphasizing the recent developments in adapting ML for addressing causal questions and policy implications (Athey, 2018; Mullainathan & Spiess, 2017; Varian, 2014; Wüthrich & Zhu, 2021).[8] Important for our context of PS estimation, machine learning is particularly instrumental when it comes to prediction (Mullainathan & Spiess, 2017). [9]

We contribute to the literature on breastfeeding in several ways. First, our study is the first (to our knowledge) to apply doubly robust estimation to the analysis of the impacts of breastfeeding on child outcomes. It is also the first in this context to apply DR and PSM techniques subject to improving the PS estimation using ML algorithms. Second, we compare the performance of alternative algorithms and offer an example of how ML may be used to improve treatment

[7] Several studies document that ML algorithms can correct for selection bias more efficiently and minimize mean square error (MSE) in predicting the treatment compared to traditional models used for PS estimation (Breiman, 2001; Friedman, 2001; Lee et al., 2010; Austin, 2012; McCaffrey et al., 2013; Ferri-García & Rueda, 2020).
[8] For example, Carmona et al. (2019) use gradient boosting to predict bank failures in the United States. In another study, Bajari et al. (2015) employed several ML algorithms to predict the demand for salty snacks using sales data and reported that ML can perform superbly in demand prediction. Khudri et al. (2023) predicted the BMI and the risks of malnutrition outcomes for Bangladeshi women of childbearing age from economic, health, and demographic features. They found that particular ML algorithms can predict the outcomes of interest more accurately and efficiently than traditional OLS and logit models.
[9] Due to the tendency of OLS to overfit a model, ML can also be considered as an alternative in the first-stage estimation of instrumental variables (Angrist & Frandsen, 2022), and can be crucial in mitigating the omitted variable bias (Athey & Imbens, 2019; Wüthrich & Zhu, 2021).

effects estimations involving PS, which can be extended to other areas of research estimating treatment effects. Third, our analysis is quite comprehensive. Using longitudinal data in a U.S. context, we evaluate several cognitive outcomes, as well as health and behavioral outcomes, at two points in time (spanning childhood and adolescence). This allows us to investigate the extent to which impacts of breastfeeding are sustained with age. We also consider possible heterogeneous treatment effects by maternal race and education, as well as breastfeeding duration, which are important for formulating targeted policies. Finally, we offer several falsification tests, as well as comparisons to results using sibling fixed effects, that provide confidence in our findings.

Our main results suggest that breastfeeding improves multiple cognitive outcomes in children between 5 and 18 years old. These findings are robust and meaningful in magnitude, with treatment effects estimated to be between 10 and 22 percent higher of a standard deviation in scores compared to that of non-breastfed peers. The estimated impacts decrease somewhat as the sample ages 10+ years but are still uniformly statistically significant. We also find significant negative associations between having been breastfed and adverse child health outcomes. Specifically, our preferred estimates from DR estimation incorporating the gradient boosting machine (GBM) algorithm suggest that breastfed children have a diminished BMI of about 10 percent, an impact that rises with age. Similarly, children who were breastfed were 8.5 percentage points less likely to be obese by the time they reached 10-18 years of age. In contrast, estimates of the relationships between breastfeeding and behavioral outcomes are more muted. We find no evidence of breastfeeding impacting a broad measure of child behavioral problems, but our estimate suggests breastfed children have a 3.6 percentage point lower likelihood of exhibiting hyperactivity by the final survey wave. Across all outcome measures, our preferred estimates are generally lower than OLS estimates, DR estimates using alternative ML algorithms, PSM estimates, and even sibling

fixed effects estimates, highlighting the importance of how one models and empirically estimates selection into treatment.

Extending our primary analysis to allow for heterogeneous treatment effects by maternal characteristics, we find that children of mothers with at least some post-high school education experience greater cognitive benefits of breastfeeding than children of less educated mothers. Our results also suggest a larger positive impact of breastfeeding on several outcomes of African American children in comparison to non-Hispanic white children, including letter word, applied problem, and a behavioral problem index, in addition to a reduced likelihood of childhood obesity among other racial/ethnic minorities. Finally, regarding breastfeeding duration, we find the benefits rise until 12 months of duration for cognitive outcomes, after which they level off. In contrast, the impacts of most health and behavioral outcomes peak at 6 months.[10]

The rest of the paper proceeds as follows. Section 2 of this study details the data sources, key variables, and estimation strategies used in the paper. Section 3 presents the results. Finally, section 4 discusses the implications of our findings and concludes.

## 2| Methods

### 2.1 | Data

#### 2.1.1 | Data source

The PSID is a leading comprehensive panel survey of a nationally representative sample comprising children, adults, and families residing in the United States. They have collected data

[10] These results support the recommendations by the USDA, the American Academy of Pediatrics, and other organizations, to engage in exclusive breastfeeding for a minimum of 6 months after birth (Meek et al., 2022; Snetselaar et al., 2021)

annually on income, wealth, employment, housing, food expenditures, health, and welfare from 1968 through 1997 and biennially following 1997. The PSID Child Development Supplement (CDS) launched in 1997 (wave 1) to investigate a wide range of children's developmental outcomes, including behavioral, cognitive, health, and other characteristics within the context of family, school, and neighborhood environments, and to gather information on children's caregivers. In the first wave of CDS, PSID formed a random sample of 3,563 children aged 0 to 12 years out of 2,394 families (McGonagle & Sastry, 2015).[11] Children and their primary caregivers (PCGs) were reinterviewed for two subsequent waves of CDS, fielded in 2002-03 (wave 2) on 2907 children aged 5–18 years, and 2007-08 (wave 3) on 1,506 children aged 10–18.[12]

We initially extract data from the first three waves of CDS and restrict the sample to children whose PCGs are the biological mother and household head or wife of the household head. Subsequently, we match the children's and their PCGs' identifications from CDS with those from the family-level and childbirth adoption and history files to access information on maternal and family characteristics and the family identification mapping file of PSID to identify siblings' information. We eliminated 287 children from the original CDS sample. Of these children, 271 were dropped as they did not live with their biological mothers during the 1997 interview, and the remaining children had missing information on their breastfeeding status. This results in a sample of 3,276 children for our analysis.

### 2.1.2 | Dependent variables

[11] Those children were born between 1984 and 1997.
[12] The children were eligible to be re-interviewed if they were 18 or younger. For further details about CDS waves, see the user guides: Hofferth et al., 1997; Mainieri, 2006; McGonagle et al., 2012.

At each wave, CDS administers subtests of the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R) to assess the reading and math skills of children over time as they progress through school.[13] CDS chose three particular subscales of WJ-R to measure achievement: the letter-word identification, the applied problems test, and the passage comprehension test.[14] The first two tests are administered among children aged three and older, while the latter is given among six years and older (Duffy & Sastry, 2014). Additionally, CDS provides test scores for broad reading combining the letter-word and passage comprehension tests.[15] Each of these four tests constitute the cognitive outcomes in our analysis. The CDS reports four types of scores for each test. We use the standardized scores adjusted for age in our analysis, which facilitates comparing children's achievements across different ages.[16]

Our study also considers four non-cognitive outcomes: body mass index (BMI) z-score[17], obesity, hyperactivity, and behavior problems index (BPI). The CDS calculates BMI z-score from the height and weight data of the children collected during the in-person interview. An obesity indicator is created following the CDC guidelines.[18] BPI measures the incidence and degree of child behavior problems in a survey setting (Peterson & Zill, 1986).[19] We use each of these outcomes from waves 2 and 3 to measure children's academic achievements, health, and behavioral progress over their early childhood and adolescence.

[13] WJ-R measures children's achievements in several dimensions of intellectual capability, including developmental status, degree of proficiency in reading and mathematics, and a group standing based on age and grade (Mainieri, 2006; Woodcock, 1997).
[14] The CDS also administered a calculation test in wave 1, only among relatively older children.
[15] Broad math score, the summation of applied problems and calculation tests, is only reported in wave 1.
[16] The CDS provides standardized scores with a mean of 100 and a standard deviation of 15 for each age group (Duffy & Sastry, 2014).
[17] For children, the Centers for Disease Control and Prevention (CDC) suggests using BMI z-score, a standardized measure of BMI using age- and gender-specific BMI distributions from the 2000 growth chart (National Center for Health Statistics, 2002).
[18] According to the 2000 growth chart, if a child's BMI is above the 95$^{th}$ percentile of the BMI distribution of the corresponding reference population, the child is considered obese (National Center for Health Statistics, 2002).
[19] Further details about the BPI scale are available in Hofferth et al. (1997).

Descriptive statistics of all outcome variables are reported in Table 1, both for the full sample and separately by whether the respondent reported ever having breastfed their child. We find an increase in the overall BMI z- score and the prevalence of obesity and hyperactivity but a decline in behavior problems index from wave 2 to 3. Comparisons in outcomes across breastfeeding status suggest that breastfed children experience better outcomes compared to their non-breastfed counterparts. Children who are breastfed receive higher test scores on all four subtests of WJ-R than their non-breastfeed peers. In contrast, breastfed children have a lower average BMI z-score and BPI score than non-breastfed children. We also find that the prevalence of obesity and hyperactivity are lower among breastfed children relative to non-breastfed ones. Of course, these differences cannot be interpreted as causal impacts of breastfeeding, as mothers and children who engaged in breastfeeding differ from those who did not.

### 2.1.3 | Independent variables

Table 2 shows the descriptive statistics for independent variables considered in our study, including the results of t-tests for differences in means between breastfed and non-breastfed sub-samples. About 44% of the sampled children were reported as ever breastfed. Nearly half of the sampled children were girls; on average, they were eight years old during the 1997 interview. Furthermore, 40% of them were firstborn. The average weight of the children at birth was approximately 7 lbs. Around 15% of the children were born small for gestational age (SGA), and 11% of them were born prematurely. According to their mothers, roughly 28% of the children were born healthier than other babies. Approximately 15% of the mothers did not complete high school, while about 34% held a college degree or higher. Over 60% of them were employed, with an average age of 27 years old. Two-thirds were married, and 40% were non-Hispanic black. About

39% reported having poor health, 45% smoked, and roughly one-fourth consumed alcohol during pregnancy. Approximately 43% of expectant mothers took part in WIC, while 22% received food stamps, and slightly over one-third were beneficiaries of Medicaid during pregnancy.

According to our descriptive statistics, mothers who breastfed their babies had higher levels of educational attainment and higher IQs, suggesting positive selection into breastfeeding. Although employed less frequently than their non-breastfeeding counterparts, these mothers tended to be in better health, smoked less, and consumed slightly more alcohol during pregnancy. Additionally, breastfeeding mothers were more likely to be married and tended to participate less in welfare enrollment programs than their non-breastfeeding peers. Families with breastfed children also had higher annual incomes.

### 2.2 | Empirical strategy

Our general empirical model is as follows:

$$Y_i = \beta_0 + \beta_1 Breastfed_i + \boldsymbol{\beta_2'} \boldsymbol{X_i} + \boldsymbol{\beta_3'} \boldsymbol{Z_i} + \varepsilon_i \quad (1)$$

where $Y_i$ represents the outcomes of children $i$. Our coefficient of interest is $\beta_1$, which captures the effects of breastfeeding on child developmental outcomes. $\boldsymbol{X_i}$ is a vector of child characteristics, including gender, age, birth order, year of birth, etc. Additionally, the vector $\boldsymbol{Z_i}$ comprises maternal, family, and community characteristics (see Table 2 for details). We derive and compare multiple estimates of $\beta_1$ using the methods described below.

#### 2.2.1| Doubly robust estimation

We use the doubly robust (DR) estimation approach to estimate the causal effect of breastfeeding on child developmental outcomes. While the DR is a selection-on-observables approach, it

minimizes bias emanating from misspecification more effectively than the traditional OLS and PSM techniques. Moreover, it offers reduced sensitivity to functional form assumptions. Both treatment (i.e., the PS model) and outcome regression models must be correctly specified to obtain unbiased estimates when they are used separately to estimate the causal effect. In contrast, the DR estimators combine these two models such that only one of them needs to be correctly specified to get an unbiased estimate (Imbens and Wooldridge, 2009; Wooldridge, 2010).

We use the augmented inverse probability weighted estimator (AIPW), a doubly robust approach, in this paper (Glynn & Quinn, 2010; Kurz, 2022; Robins & Rotnitzky, 1995; Robins et al., 1995; Robins, 2000; Scharfstein, Rotnitzky, & Robins, 1999). Although the AIPW estimator is not widely used, it contains desirable theoretical properties. It only requires researchers to: (1) specify a binary regression model for the PS, and (2) define a regression model for the outcome variable (Glynn & Quinn, 2010). The AIPW estimator uses a correction term to augment the inverse probability weighted (IPW) estimator. The term eliminates the bias if the PS model is misspecified and the outcome model is correct. Conversely, the correction term disappears if the PS model is correctly specified, and the outcome model is wrong.

If either the treatment or the outcome regression models are correctly specified, the AIPW estimator will be consistent (Ho et al., 2007; Glynn & Quinn, 2010; Scharfstein et al., 1999). When both PS and regression models are correctly specified, the AIPW estimator achieves the semiparametric efficiency bound and is more efficient than the regression adjustment or the IPW estimator (Glynn & Quinn, 2010).[20] Nevertheless, incorrect specifications of both the PS and outcome models would lead to biased estimates. Because of this possible scenario, attempting to obtain the best possible estimated PS is crucial to the robust consistency of the estimator.

[20] Additionally, Glynn & Quinn (2010) demonstrated that the AIPW estimator yields comparable or lower mean square error than the alternative estimators when both the PS and outcome models are correctly specified.

Therefore, we harness several machine learning algorithms to estimate the PS, rather than solely relying on logistic or probit regression.

The ATT estimation based on the AIPW estimator involves three steps:

1. We estimate the PS, i.e., the probability of being breastfed conditioning on observed pre-treatment characteristics using probit/logistic/supervised ML algorithms.
2. We fit a model that predicts the outcomes for the treatment group under both treatment and control conditions. The predicted outcomes are then inverse probability weighted using the PS generated from step 1 to construct a weighted average of the outcome to derive the treatment-specific predicted outcomes model (Kurz, 2022). All covariates from equation (1) are included in the outcome model.
3. Finally, we compute the average treatment (i.e., breastfeeding) effect on the treated (ATT) using the difference in predicted outcomes.[21]

The standard errors are derived by bootstrapping with 999 repetitions**,** following some other studies that used the DR estimation **(**Glynn & Quinn, 2010; Imbens, 2004; Moodie et al., 2018). The number of replications with values one less than a multiple of 100 is preferred to avoid interpolations when using the percentiles as confidence interval limits (MacKinnon, 2006).[22]

### 2.2.2 | Propensity score matching

PSM involves explicitly comparing outcomes of breastfed babies with the outcomes of individuals who were not breastfed. PSM allows us to effectively simulate an experiment by generating matched breastfed (i.e., treatment) and non-breastfed (i.e., control) samples, with both groups

[21] ATT estimates are consistent if either the treatment or the outcome model is correctly specified (Imbens & Wooldridge, 2009; Wooldridge, 2010).
[22] Moodie et al. (2018) demonstrate that a bootstrap procedure performs well even in small samples.

being similar in all other observable aspects (Rosenbaum & Rubin, 1983). After estimation of the PS, the matching estimator is non-parametric. This offers advantages over traditional regression, of which the requisite functional form and homogeneity of treatment effects assumptions are unlikely to hold (Zhao, 2008).

Our key parameter of interest from the PSM is the ATT. The ATT indicates the differences in mean outcomes between babies who were breastfed, and the same group under the hypothetical situation that they had not been breastfed. The ATT, $\delta$, is given by:

$$\begin{aligned} \delta &\equiv E[Y_1 - Y_0 | D = 1] \\ &= E\left[E\{Y_1 - Y_0 | D = 1, p(\boldsymbol{X})\}\right] \\ &= E\left[E\{Y_1 | D = 1, p(\boldsymbol{X})\} - E\{Y_0 | D = 0, p(\boldsymbol{X})\} | D = 1\right] \end{aligned} \quad (1)$$

where *D* is a dichotomous variable equal to one if the child is breastfed and zero otherwise, $Y_1$ is the outcome if treated (breastfed), and $Y_0$ is the outcome if untreated (not breastfed). $\boldsymbol{X}$ indicates a vector of covariates (pretreatment characteristics) included in our models and $p(\boldsymbol{X})$, the PS , is the conditional probability of receiving a treatment (i.e., being breastfed) given observed characteristics. To interpret estimates of the ATT ($\delta$) as causal, it must be assumed that the PS includes all relevant variables, or that the effect of unobservable characteristics on the PS is the same as that of observable characteristics.

To generate estimates of $\delta$, we first apply a (single) nearest-neighbor matching procedure to match breastfed children to comparable children who were not breastfed.[23] For all our estimation involving PS, to help ensure similar treated and control units we impose a common support restriction on both treatment and control groups. Treated and control units whose estimated PS lies

[23] We employ the "Matchit" package (Ho et al., 2011), matching with replacement.

outside the intersection of the range of PS of treated units and the range of PS of the control units are discarded.[24]

We further evaluate the quality of the matching process by checking for covariate balance resulting from the matched sample. Appendix Table A1 reports the mean values of the covariates used in the analysis by treatment and control groups after performing nearest-neighbors matching for each test wave.[25] Our overall matching results corroborate that our treatment and control groups are quite similar after matching, with insignificant differences in all variables except for only one in the PSM for cognitive outcomes at a 10% level.

### 2.2.3 | Propensity score adjustment using machine learning algorithms

The traditional PS estimation methods used in the economics literature are logistic and probit regression models. However, these methods may overestimate the treatment effects if they fail to remove non-observable bias and minimize MSE effectively in predicting the probability of assignment to treatment (Lee et al., 2010; Ferri-García & Rueda, 2020). The efficacy of the PSM at removing non-observable bias depends primarily on the selection mechanism, covariates chosen, and data dimensionality. Another drawback of the logistic model is that it assumes the log-odds risks are linearly associated with the covariates (Agresti, 2012). This assumption might fail to hold while testing the hypothesis that breastfeeding affects a child's subsequent developmental outcomes, with a large sample and many covariates.

[24] 3.85% of the observations were dropped due not meeting the common support requirements. The resulting range of propensity score estimates for both groups from PSM with gradient boosting algorithm is shown in Appendix Figure A1. Substantial representation of both groups is present across the full range of common support.
[25] The covariate balance analysis in Appendix Table A1 is derived from nearest-neighbors matching using GBM-based PS (described in Section 2.2.3). Covariate balance derived from alternative PS estimates is comparable.

Black-box (e.g., neural network and bagging/boosting algorithms) and interpretable machine learning algorithms such as classification and regression trees are recommended as potential replacements subject to the available covariates and the complexity of the relationships impacting selection (Breiman, 2001; Friedman, 2001; Lee et al., 2010). Athey and Imbens (2017) discuss the role of machine learning methods in estimating PS. These methods may improve upon traditional PS estimation based on logistic regression, which performs poorly under model misspecification. McCaffrey et al. (2004) emphasize that while traditional PS methods can remove confounders, many covariates could lead to an inaccurate estimation of the PS. They apply boosting to overcome this issue and show that PS weights computed based on boosting removed pretreatment group differences and obtained a better estimate for the treatment.

Boosting, discussed by Athey and Imbens (2019) as a way to improve the performance of supervised leaning methods, involves building regression trees sequentially by placing weights on observations that incorrectly classify the outcome. This can reduce bias and yield more stable weights than logit, probit, and multinomial models at the cost of a minimal rise in variance (Austin, 2012; McCaffrey et al., 2013). Gradient boosting machine (GBM) is the upgraded version of boosting, and it allows for both categorical and continuous outcomes. In GBM, trees are sequentially built to minimize prediction error measures such as the mean squared error (Friedman, 2001).[26] Additionally, GBM models perform well with rare outcomes and a smaller sample size. For example, Ayaru et al. (2015) predicted acute lower gastrointestinal bleeding using 300 observations and demonstrated that GBM predicted the outcome with a higher accuracy and outperformed multiple logistic regression-based models.

[26] See Appendix B for methodological details of the GBM algorithm.

While our ultimate aim is to estimate treatment effects of breastfeeding by incorporating the best performing algorithm, an additional objective of our study is to evaluate and compare the performance of GBM and several widely used ML algorithms (along with traditional logit and probit models) in predicting the probability of being breastfed. Specifically, we consider logit, probit, classification and regression trees (CART), random forest (RF), gradient boosting machine (GBM), neural network (NN), and generalized additive model (GAM) for estimating the PS. Tables A2 and A3 report the specifications of the SML algorithms used for estimating the PS in the DR and PSM estimations, respectively.[27] We calculate the $R^2$, mean squared error (MSE), bias, mean absolute error (MAE), and mean absolute percentage error (MAPE) from the out-of-sample to evaluate the prediction error and performance of the algorithms used in the PS estimation. Results from all these performance measures are shown in Figure 1 and indicate that the GBM outperforms other algorithms in predicting the probability of being breastfed most accurately, as evidenced by its highest $R^2$ value, and lowest MSE, bias, MAE, and MAPE values.[28] Specifically, the MSE value from the GBM is about 30% and 31% lower than that of the logit and probit models, respectively. Additionally, the $R^2$ value of GBM exceeds that of the logit and probit models by about 26 and 28 percentage points, respectively. Therefore, we selected the ATT estimation utilizing GBM-based PS as our primary estimation technique for treatment effects.

[27] To avoid overfitting, the supervised machine learning algorithms are regularized by imposing a penalty on the model when additional variables are included. To what extent a model requires to be regularized depends on cross-validation (Daoud et al., 2019). Typically, the SML algorithms split the data into an in-sample (also known as the training dataset) portion that are used to train the model and an out-of-sample (also known as the test dataset) portion to evaluate its prediction performance. We employ 80% of the data as the training dataset, while the remaining 20% as the test dataset. Next, we apply 10-fold cross-validation to the training set to optimize out-of-sample prediction (Mullainathan & Spiess, 2017). This involves randomly splitting the training data set into ten equal sizes (or folds), with nine folds randomly chosen to train the model, and the remaining fold used to assess its performance. These is performed *ten* times with each fold alternatively acting as the testing set, and the collective predictive performance is used to determine the optimal parameters of the model.

[28] The GBM algorithm surpasses other algorithms across all performance evaluation measures other than the MAE in estimating the PS used in the PSM technique. Results are available upon request.

### 2.2.4 | Sibling fixed effects

One of the limitations associated with PSM is that it does not rule out unobserved heterogeneity arising from genetic and family-related factors that could impact estimates. Sibling fixed-effects models provide a way to control for unobserved, time-invariant hereditary and household factors (e.g., home environment, genetics, mother's characteristics) that are common across siblings. The structure of the data allows us to observe siblings from the same mother, who each may or may not have been breastfed. The sibling fixed-effects model exploits within-mother variation, i.e., sibling comparisons, to estimate the effect of breastfeeding on a child's developmental outcomes. A significant drawback of this approach is that it focuses solely on households with more than one child. Also, a mother's decision to breastfeed is likely to apply to both her babies, which further restricts the identifying variation. Further, siblings' differential abilities to breastfeed can be correlated with ensuing developmental outcomes. Despite these drawbacks, for comparison to our primary PSM estimates we estimate the following fixed-effects models using the data on 1062 pairs of siblings included by the CDS:[29]

$$Y_{ij} = \beta_0 + \beta_1 Breasted_{ij} + \boldsymbol{\beta_2'} \boldsymbol{X_i} + \boldsymbol{\tau_j} + \varepsilon_{ij} \quad (2)$$

where $Y_{ij}$ represents the outcomes of children *i* at family *j*. Our coefficient of interest is $\beta_1$, which captures the relationship between breastfeeding and child developmental outcomes. The vector $\boldsymbol{X_i}$ comprises the set of sibling-varying controls that are used in the PSM and $\boldsymbol{\beta_2}$ indicates the vector of corresponding coefficients. $\boldsymbol{\tau_j}$ is a vector of sibling fixed effects which controls for all factors

[29] We discard the mother with only one child from our sample, leaving a total sample size of 2124 for sibling fixed effects analysis.

that are common to both siblings, ruling out the need to observe and measure potentially critical confounders (Rees and Sabia, 2009).

## 3| Results

In this section we report and discuss estimates of treatment effects of ever having been breastfed (i.e., breastfeeding initiation) on child developmental outcomes. We focus primarily on results from DR and PSM techniques in which PS are estimated using GBM algorithm. After presenting these results, we turn to sibling fixed effects estimates. Following this, we consider possible heterogeneity in treatment effects along several dimensions: breastfeeding duration, children's age, and maternal race and education.

### 3.1| Main estimates

Table 3 presents baseline estimates of the relationships between breastfeeding initiation and cognitive and non-cognitive outcomes using and DR estimation techniques. For comparison purposes, column 1 displays the estimates from OLS with the same covariates, whereas columns 2 and 3 report the ATT estimates from the PSM and DR techniques in which PS are estimated using the GBM algorithm.

Our analysis shows that breastfeeding is significantly associated with improved cognitive outcomes (i.e., higher test scores) in both waves 2 (when children were aged 5 to 18 years) and wave 3 (when children were aged 10 to 18 years). This finding aligns with previous studies that identified a positive link between breastfeeding and cognitive test scores (Belfield & Kelley, 2012; Borra et al., 2012; Fitzsimons & Vera-Hernández, 2022; Rees & Sabia, 2009). This result is robust

across all four test scores and across all specifications. In terms of magnitude, the estimated impacts are modest but meaningful. For instance, the estimated ATT for letter word test scores obtained from the DR estimation in wave 2 is 3.093, relative to a mean score in the sample of about 100. Given the standard deviation in this test score, we find that breastfed children performed about 17 percent of a standard deviation higher than non-breastfed children. Across all test scores, breastfed children performed better than those who were not, with an improvement of about 10%-22% of a standard deviation. The DR approach (with GBM), which may perform better than OLS and PSM in dealing with selection into breastfeeding, results in relatively lower effect sizes among corresponding estimates while controlling for the same set of covariates. The estimated impacts on cognitive outcomes diminish somewhat as children age from wave 2 to wave 3 (five years later), although they remain statistically significant. The most persistent treatment effect is observed for passage comprehension, while the impact on letter word score falls dramatically and the significance level is reduced to 10%.

Turning to non-cognitive outcomes, each specification suggests that being breastfed is associated with a decline in a child's BMI z-score. The estimates range from -0.099 to -0.144 in wave 2 and increase to between -0.254 to -0.355 when children are older in wave 3.[30] In line with these findings, we also find that breastfed children are less likely to be obese, with estimates ranging from 3.7 to 6.2 percentage points in wave 2 and 8.5 to 10.4 percentage points in wave 3. These results are consistent with the study by Metzger & McDade (2010), which found that breastfed children had lower BMI scores than their non-breastfed siblings. They also found that breastfed siblings were less likely to be overweight or obese. We also observe breastfeeding to be associated with a lower risk of being hyperactive by 2.5 to 2.6 percentage points at earlier ages

[30] Gibson et al. (2017) found statistically significant effect of breastfeeding on BMI, specifically in older children.

(wave 2), though this result is statistically insignificant for the PSM specification and marginally significant for the OLS and DR specifications. Conversely, breastfeeding appears to be statistically significantly linked with a 3.6-6.7 percentage points decrease in hyperactivity risk as children grow older. These findings are consistent with previous studies (Soled et al., 2021; Brasfield, Goulding, & Kancherla, 2021). However, our results do not provide evidence of a significant relationship between breastfeeding and behavior problems, either at an early age or later. This supports the findings of Borra et al. (2012) and Belfield & Kelley (2010).

It is worth comparing our primary results to those obtained using traditional models of PS estimation (i.e., logit or probit) as well as alternative SML methods of PS estimation. These comparisons of DR estimates are displayed in Appendix Table A4. Two broad characterizations are notable from these comparisons. First, the results from various models suggest that the general findings from our primary estimates are robust. Specifically, breastfeeding is linked to higher test scores and improved cognitive outcomes in both early and later childhood stages, with the effect diminishing over time. As for non-cognitive outcomes, all the algorithms indicate that breastfeeding significantly reduces a child's BMI z-score, as well as the risk of obesity in both waves. Most algorithms suggest that breastfeeding is associated with a lower risk of hyperactivity, with a pronounced effect in the latter wave. In contrast, behavior problems seem to have no significant association with breastfeeding across most of the algorithms. Second, the GBM-based DR approach yields lower point estimates of ATT values compared to other algorithms and consistently provides lower estimates than traditional logistic and probit models.[31] We also

[31] Out of all sixteen outcomes, we found significant differences (based on a t-test) in ATT values at a 10% level between GBM and logistic or probit models in letter word, applied problem, and passage comprehension scores in Wave 2. Even though only a few outcomes showed significant differences, the results reported in Table A4 collectively reinforce the argument that machine learning algorithms can yield a lower ATT value than the traditional approaches subject to reducing bias.

estimate the PS using the same set of SML algorithms for PSM estimation, and the results again suggest that the GBM-based ATT estimates are lower for most of the outcomes, particularly for cognitive scores (Appendix Table A5). This finding suggests the possibility of overestimation of treatment effects due to the prevailing bias resulting from an error in calculating the PS using alternative methods.

### 3.2 | Sibling fixed effects estimates

To further evaluate the robustness of our primary results using PSM and DR estimation, and to consider the potential impact of unobserved factors biasing our results, we present results from sibling fixed effects models. If there exists selection into breastfeeding initiation on the basis of any family background, home environment or genetic characteristics that are common across siblings and may also independently influence outcomes, the use of sibling effects will eliminate their role in the estimation. However, a downside of this approach is that we must confine our sample to multiple-child households. Furthermore, ATT estimates are identified solely on sets of siblings for whom there exists variation in treatment status. These caveats aside, the findings (presented in Table 4) are largely consistent with the OLS, PSM, and DR estimates, with cognitive outcomes being positively impacted by breastfeeding and, with the exception of behavior problems, non-cognitive outcomes being significantly negatively impacted. In comparing the magnitudes of our estimates, the DR estimates are almost uniformly lower than those of the sibling fixed effects model, except for the hyperactivity outcome in Wave 2. This is notable given that breastfeeding generally exhibits positive selection on observables, which is typically assumed for unobservables as well. So, while the sample restrictions on sibling fixed effects estimation make

the comparisons imperfect, our results from the sibling fixed effects model serve as a robustness check and provide confidence that the role of omitted variable bias is limited.

### 3.3| Heterogeneous treatment effects

To summarize our primary estimation results so far, we find that breastfeeding improves cognitive test scores, and the results persist across different techniques. In contrast, our results based on non-cognitive outcomes are mixed. Having been breastfed is negatively associated with both BMI and obesity, but the association with behavioral outcomes is not consistent. We now turn to stratifying the sample by breastfeeding duration, children's age, maternal race, and maternal education to investigate if our estimated results vary across these dimensions.

#### 3.3.1| Heterogeneity of results by breastfeeding duration

Thus far we have considered the impacts of ever having been breastfed. We now analyze how varying durations of breastfeeding are associated with child developmental outcomes (Table 5).[32] We focus on four different periods of breastfeeding: breastfeeding for at most 3 months, at most 6 months, 6+ to 12 months, and more than 12 months.[33] We compare each duration category with reference to never having been breastfed. Our findings show that all durations of breastfeeding up to 12 months positively and significantly impact cognitive outcomes. Although the positive effects

[32] Results from the PSM (omitted for brevity) are quantitatively similar. They are available upon request.

[33] Rather than a continuous measure, we focus on these categories for a few reasons. First, the retrospective nature of the breastfeeding data makes measurement error in precise duration likely. Second, the duration of breastfeeding is truncated at 24 months for those who were breastfed for a longer period ($n$=27). Finally, these category cutoffs correspond to common recommendations of breastfeeding duration. According to the National Health and Nutrition Examination Survey (NHANES) during 1988-1994, the prevalence of breastfeeding at 6 and 12 months in wave 3 were 22.4% and 8.9%, respectively (Li et al., 2002), versus 23.56% and 10.4% from our study using the PSID data, providing reassurance of the consistency between studies.

on applied problem and passage comprehension test scores are only substantial in Wave 2 for the 12+ month duration, the effects on other cognitive outcomes remain positive but level off.[34] We observe a greater benefit of breastfeeding among those who were breastfed for 6+ to 12 months compared to those who were breastfed for a maximum of 3 months. Also, children who were breastfed for 6+ to 12 months appear to have a larger effect on letter word, applied problem, passage comprehension (in wave 3), and broad reading than those breastfed for a maximum of 6 months. Overall, these results suggest a positive but diminishing impact of breastfeeding, and that breastfeeding beyond 12 months is unlikely to offer additional cognitive benefits to children. The maximum impacts of breastfeeding on cognitive outcomes are estimated to occur between 6 and 12 months of breastfeeding for letter word and applied problem scores, each estimated to increase scores by about 0.3 standard deviations compared to never having been breastfed.

Our results indicate that breastfeeding for 3 months or 6 months is linked to a decrease in the child's BMI z-score and the risks of being obese and hyperactive in both waves, in contrast to those who were never breastfed. Furthermore, children who were breastfed for 6+ to 12 months also had a decrease in BMI z-score and lower risks of being obese and hyperactive in Wave 3 than their non-breastfeeding counterparts. However, our findings indicate that breastfeeding does not affect behavioral problems, regardless of duration. Overall, our results suggest that the benefits of breastfeeding rise until 6 months of breastfeeding and subsequently flatten off for non-cognitive outcomes.

### 3.3.2| Heterogeneity of results by child's age

[34] It should be noted that the sample size of this group is relatively small (3.5 percent of the overall sample).

In this exercise, we limit our sample to children who were aged 7 or younger during the 1997 interview (wave 1) and rerun our primary analysis. The purpose of this analysis is twofold. Firstly, focusing on the earlier wave helps to minimize the possibility of recall bias.[35] Secondly, we wish to examine whether there is any variation in our estimated treatment effects across waves among the same children. This provides an alternative way to investigate whether the impact of breastfeeding on developmental outcomes changes with children's age. ATT estimates from the DR estimation are presented in Table 6. According to our estimates, breastfeeding has a positive association with test scores for all four cognitive outcomes in both waves. Nonetheless, we observe more negligible treatment effects in the last wave when the children get older, implying that the impact of breastfeeding on cognitive abilities may decline over time. However, these declines are not statistically significant.

In contrast, our findings indicate that breastfeeding has a notably greater impact on some noncognitive outcomes, particularly BMI and obesity, in the third wave when children are older versus wave 2. Breastfeeding is found to be significantly associated with a lower risk of hyperactivity in both waves. In contrast, we observe no significant relationship between breastfeeding and behavior problems across both waves. All of these results validate the baseline estimates reported in Table 3.

### 3.3.3| Heterogeneity of results by maternal race

[35] As PSID collects retrospective cohort data, our results could be subject to some recall bias owing to the long lag between birth and the survey, causing a potential downward bias in estimates. To address this issue, we first investigate those who reported breastfeeding in both the 1997 and 2002 waves and see if their responses are consistent. We find a correlation of 0.86, thus suggesting a relatively low recall bias.

Despite the increase in the overall breastfeeding rates in the United States over the past decade, racial/ethnic disparities still persist. Black women and some other racial and ethnic minority groups continue to have lower breastfeeding rates and are far from meeting the Healthy People 2020 goals, as stated in studies by Jones et al. (2015) and Chiang et al. (2021). To investigate whether estimates differ across maternal race/ethnicity, the baseline model was re-estimated by stratifying the sample based on maternal race/ethnicity, which includes non-Hispanic white (*n*=1534), non-Hispanic black (*n*=1305), and Other race/ethnicity[36] (*n*=424). Table 7 presents the ATT values from the DR estimation for each group.[37] Our results, in general, suggest that children with minority mothers have significantly higher cognitive test scores compared to children with non-Hispanic white mothers. In particular, non-Hispanic black children observe the most considerable impact of breastfeeding on letter words, applied problems, and broad reading scores. For these children in wave 2, the estimated impacts of having been breastfed on letter word score and applied problem score are about 0.41 and 0.48 standard deviations higher, respectively. These estimates are over twice as high as the estimates for the non-Hispanic white sample, with both differences statistically significant at the 10 percent level. With the exception of applied problem in both waves and passage comprehension in wave 3, the estimated treatment effects are also higher for the "other race" sample than they are for children of non-Hispanic white mothers, though the differences are not statistically significant.

We also observe some differences by maternal race for non-cognitive outcomes. Specifically, we find significant associations of breastfeeding with child obesity and BMI only for the non-Hispanic white and other race samples, and not for the non-Hispanic black sample. The estimated reduction in the likelihood of obesity is especially large for the "Other race" sample. In

[36] Other race/ethnicity refers to Hispanic, Asian, and other (combined due to lower sample sizes).
[37] Results from the PSM are available upon request.

contrast, breastfed children with non-Hispanic black mothers appear to have the largest reduction in hyperactivity in both waves. Similarly, the association between having been breastfed and the likelihood of behavioral problems is (highly) statistically significant only for this group, reflecting an approximate 0.16 standard deviation reduction in this outcome.

Other studies tend to not separate their analysis by race across ages. Therefore, comparing our results with prior studies is difficult. Nevertheless, previous studies have investigated the benefits of breastfeeding for children below the poverty line (Belfield & Kelly, 2010) and socially disadvantaged groups (McCrory & Layte, 2011) and have found larger benefits for underprivileged groups.

### 3.3.4| Heterogeneity of results by maternal education

We also run a heterogeneity analysis by stratifying the sample by maternal education (Table 8). We find that children of mothers with at least some post-high school education (50% of the sample) experience relatively greater benefits from breastfeeding. Although this finding contrasts with results from Borra et al. (2012) and Fitzsimons & Vera-Hernández (2022), both UK-based studies, they corroborate the results of Gibson-Davis & Brooks-Gunn (2006), who only found positive breastfeeding effects for mothers with at least some post-secondary education in the United States. In particular, while our estimated impacts on cognitive outcomes are positive and significant for children of mothers with high school or lower educations only in wave 1, the estimates are uniformly higher and statistically significant among the children of higher educated mothers in both waves. Furthermore, the results from t-tests suggest that the differences across education groups are largely statistically significant. Regarding non-cognitive outcomes, we notice that breastfed children of mothers with some college degrees experience a significant decrease in BMI

and lower risks of obesity and hyperactivity in both waves. Interestingly, we find larger health-related (BMI and obesity) effects in wave 3 among the children of lesser educated women than we do among the children of mothers with some college degrees. In contrast to our cognitive results, we find no statistically significant differences in ATT estimates by maternal education for the health and behavioral outcomes.

### 3.4| Robustness checks

#### 3.4.1 | Excluding low birthweight and pre-mature births

The relationships between breastfeeding and children's developmental outcomes may be influenced by unobservable circumstances related to a child's birth, such as prematurity. Furthermore, infants with poor health may face challenges latching and are less likely to be breastfed (U.S. Department of Health and Human Services, 2011). This poses a problem of structural endogeneity that could potentially bias our results. To ensure the robustness of our findings, we remove low-birthweight infants (i.e., those weighing less than 5.51 lbs. or 2.5 kg during birth) and premature babies (i.e., infants born before 37 weeks), as these groups may differ systematically from the rest of our sample.

Table 9 reports the effects of breastfeeding on child developmental outcomes after excluding low birth weight and premature babies. Overall, the results are very similar to our main estimates and suggest that breastfeeding significantly improves all test scores in early childhood. Although the effects tend to diminish over time, they are sustained for all cognitive outcomes at a later age, except for letter word scores. This provides further evidence that breastfeeding is a valuable practice that can help promote the cognitive development of children.

The results further show that breastfeeding is linked with a reduction in the child's BMI z-score and a lower likelihood of obesity at both points in time. In contrast, breastfeeding appears to affect hyperactivity only in Wave 3, while its association with behavior problems turns out to be statistically insignificant. Overall, the estimates for non-cognitive outcomes are similar to those in the baseline estimates. This suggests our primary findings are not driven by endogeneity resulting from low birth weight and prematurity.

### 3.4.2 | Falsification exercise

In this section, we perform a falsification exercise to assess whether unobservables correlated with breastfeeding could drive our baseline estimates. This strategy requires identifying a set of outcomes that could be linked to these unobservables but are unlikely to be directly affected by breastfeeding. Especially if outcomes are measured pre-treatment, i.e., before breastfeeding occurs, they should be unaffected by breastfeeding unless there exists some selection bias that the PSM technique does not account for. The conditional independence assumption is more reasonable if the results suggest that breastfeeding does not impact the placebo outcome. If variables chosen in the falsification exercise are closely associated with the outcome of interest, the exercise has more acceptability (Imbens & Wooldridge, 2009). We chose five placebo outcomes in this exercise: child's birth weight, whether the mother smoked six months before pregnancy, whether the mother smoked during pregnancy, whether the mother drank alcohol 12 months before pregnancy, and whether the mother drank alcohol during pregnancy. Birth weight is closely associated with both cognitive and noncognitive outcomes (Imbens & Wooldridge, 2009) and can, therefore, serve as a crucial pre-treatment proxy outcome. The other variables may indicate

maternal attitudes towards their children or parenting and could be linked to unobservables affecting children's outcomes.

Table 10 displays results from the falsification exercise using sibling fixed effects, PSM, and DR techniques. All the specifications indicate that the relationships between the placebo outcomes and breastfeeding fall short of statistical significance at conventional levels, providing no evidence of selection bias.

## 4| Discussion and conclusion

This study is the first to incorporate doubly robust estimation and machine learning methods in estimating the impact of breastfeeding on child outcomes. Our comprehensive analysis, including baseline estimates and heterogeneity analysis by varying durations of breastfeeding, children's age, maternal race, and education, suggests that breastfeeding is significantly linked to multiple improved cognitive outcomes in early childhood. Our baseline estimates indicate that breastfed children (between 5 and 18 years of age) outperform those who are not breastfed by about 10%–22% of a standard deviation across cognitive test scores. These effects largely persist as the sample ages, with the exception of letter word score, for which the estimate drops in magnitude by about 40 percent. Further, we find that having been breastfed is negatively associated with BMI and obesity among children, especially among older children. In wave 3, among children aged 10-18, our results suggest breastfeeding reduced BMI by about 14 percent of a standard deviation, and the likelihood of obesity decreased by 8.5 percentage points. Finally, we find weak evidence that hyperactivity was reduced by a small degree among breastfed children, but no evidence that breastfeeding impacted a broad behavioral problems index.

Though meaningful in magnitude, these estimated treatment effects resulting from our

preferred (bias and MSE-minimizing) algorithm are generally lower than those found using OLS, matching estimators using alternative estimation of PS, and sibling fixed effects. Our estimates on cognitive scores are also lower than some notable estimates in the literature derived from experimental or natural experimental methods. For example, Fitzsimons & Vera-Hernández (2022) find impacts of breastfeeding on cognitive outcomes of about 0.5 standard deviations – over twice the magnitude of our estimates. One possible reason for this discrepancy could be the difference in settings, with their sample drawn from vaginal births in public hospitals in the United Kingdom. Also, their estimates represent a local average treatment effect, with identification driven solely by those mothers who only chose to breastfeed due to increased breastfeeding support services available in the hospital upon giving birth. Since these mothers may be less likely to make other early investments in their children's development, the relative impact of breastfeeding could be quite large.

Our other findings offer policymakers additional insight into the nuanced relationships between children's health capital investments and multiple developmental outcomes. A particularly noteworthy finding of our study is that the benefits of breastfeeding rise until 12 months of breastfeeding duration for cognitive outcomes (and 6 months for health-related outcomes) and flatten off subsequently. This result is in line with some previous studies (Binns et al., 2016; Horta et al., 2015; Kramer et al., 2008a,b), which suggest that the benefits of breastfeeding may rise beyond 6 months. Our results also support the recommendations by the USDA, the American Academy of Pediatrics, and other organizations, which encourage engagement in exclusive breastfeeding for a minimum of 6 months after birth, with some continued breastfeeding after that age (Meek et al., 2022; Snetselaar, et al., 2021).

Our analysis also suggests that there are some heterogenous impacts of breastfeeding across

demographic groups. We find that children of mothers with at least some post-high school education experience greater benefits of breastfeeding than children of less educated mothers**.** There is a large socioeconomic gradient in mothers' breastfeeding activity in the United States, with 75.8 percent of college graduates reporting breastfeeding at 3 months, and only 37.8 percent of those with high school or less reporting the same (Diaz et al., 2023). While this differential is in line with our relative magnitudes of estimated cognitive benefits of breastfeeding, we instead find larger health-related (BMI and obesity) effects among lesser educated women than we do among those with college degrees. We also observe a few differentials in estimated treatment effects by race, with children of African Americans benefitting more on letter word score, applied problem score, and the behavioral problem index, in comparison to children of non-Hispanic, white mothers. Especially given the low rates of breastfeeding among African Americans (36.3 percent), our results suggest breastfeeding informational or support policies targeting this group of women would be particularly beneficial.

Methodologically, our analysis highlights the role that ML methods can have in improving PS estimation. Reliable PSM estimators require accurate estimation of the PS, and even doubly robust methods benefit from improved PS estimation. ML methods excel at such prediction tasks. Accordingly, we find that such methods – in particular the GBM algorithm – result in lower bias and mean squared error than traditional methods of estimating the PS. Future researchers should consider using ML, and comparing the relative performance of particular algorithms, in other contexts involving PS estimation.

Our analysis is not without limitations. Most importantly, since our analysis inherently relies on observable variables to balance treatment and control groups, we cannot rule out the possibility that some relevant factors, related to both child outcomes and a mother's choice to

breastfeed, have been omitted. However, the results from falsification exercises and comparable or lower DR estimates than those of sibling fixed effects provide some confidence that the role of omitted variables is limited. Second, due to data limitations, we were only able to examine the impact of ever having been breastfed (as well as different durations of breastfeeding), versus the impact of exclusive breastfeeding. Some prior research has found exclusive breastfeeding to result in stronger outcomes (Del Bono & Rabe, 2012), while others have found no difference (Borra et al., 2012).

**References**


Adair, L. S., & Guilkey, D. K. (1997). Age-specific determinants of stunting in Filipino children. *The Journal of nutrition*, *127*(2), 314-320.

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons.

Alexander, G. R., Himes, J. H., Kaufman, R. B., Mor, J., & Kogan, M. (1996). A United States national reference for fetal growth. *Obstetrics & Gynecology*, *87*(2), 163-168.

Almond, D., Currie, J., & Duque, V. (2018). Child circumstances and adult outcomes: Act II. *Journal of Economic Literature*, 56(40), 1360-1446.

Angrist, J. D., & Frandsen, B. (2022). Machine labor. *Journal of Labor Economics*, *40*(S1), S97-S140.

Athey, S. (2018). The impact of machine learning on economics. In *The Economics of Artificial Intelligence: An Agenda* (pp. 507–547). University of Chicago Press.

Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. *Journal of Economic Perspectives*, *31*(2), 3–32.

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics*, *pp. 11*, 685–725.

Austin, P. C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behavior research*, *47*(1), 115-135.

Ayaru, L., Ypsilantis, P. P., Nanapragasam, A., Choi, R. C. H., Thillanathan, A., Min-Ho, L., & Montana, G. (2015). Prediction of outcome in acute lower gastrointestinal bleeding using gradient boosting. PLoS One, 10(7), e0132485.

Bajari, P., Nekipelov, D., Ryan, S. P., & Yang, M. (2015). Machine learning methods for demand estimation. *American Economic Review*, *105*(5), 481-485.

Baker, M., & Milligan, K. (2008). Maternal employment, breastfeeding, and health: Evidence from maternity leave mandates. *Journal of Health Economics*, *27*(4), 871-887.

Belfield, C. R., & Kelly, I. R. (2012). The benefits of breast feeding across the early years of childhood. *Journal of Human Capital*, *6*(3), 251-277.

Bigelow, A. E., & Power, M. (2020). Mother-infant skin-to-skin contact: short- and long-term effects for mothers and their children born full-term. *Frontiers in Psychology*, 11: 1921.

Binns, C., Lee, M., & Low, W. Y. (2016). The long-term public health benefits of breastfeeding. *Asia Pacific Journal of Public Health*, *28*(1), 7-14.

Borra, C., Iacovou, M., & Sevilla, A. (2012). The effect of breastfeeding on children's cognitive and noncognitive development. *Labour Economics*, *19*(4), 496-515.

Brasfield, J., Goulding, S. M., & Kancherla, V. (2021). Duration of breast feeding and attention-deficit/hyperactivity disorder in United States preschool-aged children. *Research in Developmental Disabilities*, *115*, 103995.

Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

Brennan, L., McDonald, J., & Shlomowitz, R. (2004). Infant feeding practices and chronic child malnutrition in the Indian states of Karnataka and Uttar Pradesh. *Economics & Human Biology*, *2*(1), 139-158.

Caldwell, B. M., & Bradley, R. H. (1984). Home observation for measurement of the environment. Little Rock, AR.

Carmona, P., Climent, F., & Momparler, A. (2019). Predicting failure in the US banking sector: An extreme gradient boosting approach. *International Review of Economics & Finance*, *61*, 304-323.

Case, A., & Paxson, C. (2008). Stature and status: Height, ability, and labor market outcomes. *Journal of Political Economy*, *116*(3), 499-532.

Chiang, K. V., Li, R., Anstey, E. H., & Perrine, C. G. (2021). Racial and ethnic disparities in breastfeeding initiation─ United States, 2019. *Morbidity and Mortality Weekly Report*, *70*(21), 769.

Cunha, F., Heckman, J. J., & Schennach, S. M. (2010). Estimating the technology of cognitive and noncognitive skill formation. *Econometrica*, *78*(3), 883-931.

Daoud, A., Kim, R., & Subramanian, S. V. (2019). Predicting women's height from their socioeconomic status: A machine learning approach. *Social Science & Medicine*, *238*, 112486.

Dehejia, R. H., & Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, *84*(1), 151-161.

Del Bono, E., & Rabe, B. (2012). *Breastfeeding and child cognitive outcomes: Evidence from a hospital-based breastfeeding support policy* (No. 2012-29). ISER Working Paper Series.

Diaz, L. E., Yee, L. M., & Feinglass, J. (2023). Rates of breastfeeding initiation and duration in the United States: data insights from the 2016-2019 Pregnancy Risk Assessment Monitoring System. *Frontiers in Public Health*, 11, 1256432.

Duffy, D., & Sastry, N. (2014). Achievement tests in the panel study of income dynamics child development supplement. PSID Technical Series Papers, 14-02.

Dyson, L., Renfrew, M., McFadden, A., McCormick, F., Herbert, G., & Thomas, J. (2006). Promotion of breastfeeding initiation and duration. *Evidence into practice briefing. London: NICE*.

Evenhouse, E., & Reilly, S. (2005). Improved estimates of the benefits of breastfeeding using sibling comparisons to reduce selection bias. *Health services research*, *40*(6p1), 1781-1802.

Ferri-García, R., & Rueda, M. D. M. (2020). Propensity score adjustment using machine learning classification algorithms to control selection bias in online surveys. *PloS one*, *15*(4), e0231500.

Fitzsimons, E., & Vera-Hernández, M. (2022). Breastfeeding and child development. *American Economic Journal: Applied Economics*, *14*(3), 329-66.

Francesconi, M., & Heckman, J. J. (2016). Child development and parental investment: Introduction. *The Economic Journal*, *126*(596), F1-F27.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, *29*(5), 1189-1232.

Gibson, L. A., Hernández Alava, M., Kelly, M. P., & Campbell, M. J. (2017). The effects of breastfeeding on childhood BMI: a propensity score matching approach. *Journal of Public Health*, *39*(4), e152-e160.

Gibson-Davis, C. M., & Brooks-Gunn, J. (2006). Breastfeeding and verbal ability of 3-year-olds in a multicity sample. *Pediatrics*, *118*(5), e1444-e1451.

Glynn, A. N., & Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political analysis*, *18*(1), 36-56.

Haines, M. R., & Kintner, H. J. (2008). Can breast feeding help you in later life? Evidence from German military heights in the early 20th century. *Economics & Human Biology*, *6*(3), 420-430.

Heckman, J. J., & Mosso, S. (2014). The economics of human development and social mobility. *Annu. Rev. Econ.*, *6*(1), 689-733.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political analysis*, *15*(3), 199-236.

Ho, D., Imai, K., King, G., & Stuart, E. (2011). MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, *42*(8), 1-28.

Hofferth, S., Davis-Kean, P. E., Davis, J., & Finkelstein, J. (1997). The child development supplement to the Panel Study of Income Dynamics: 1997 user guide. Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

Horta, B. L., Loret de Mola, C., & Victora, C. G. (2015). Breastfeeding and intelligence: a systematic review and meta-analysis. *Acta paediatrica*, *104*, 14-19.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, *86*(1), 4-29.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, *47*(1), 5-86.

Ip, S., Chung, M., Raman, G., Chew, P., Magula, N., DeVine, D., ... & Lau, J. (2007). Breastfeeding and maternal and infant health outcomes in developed countries. *Evidence report/technology assessment*, (153), 1-186.

Jones, K. M., Power, M. L., Queenan, J. T., & Schulkin, J. (2015). Racial and ethnic disparities in breastfeeding. *Breastfeeding Medicine*, *10*(4), 186-196.

Khudri, M. M., Rhee, K. K., Hasan, M. S., & Ahsan, K. Z. (2023). Predicting nutritional status for women of childbearing age from their economic, health, and demographic features: A supervised machine learning approach. *Plos One*, *18*(5), e0277738.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The quarterly journal of economics*, *133*(1), 237-293.

Kramer, M. S., Aboud, F., Mironova, E., Vanilovich, I., Platt, R. W., Matush, L., ... & Promotion of Breastfeeding Intervention Trial (PROBIT) Study Group. (2008a). Breastfeeding and child cognitive development: new evidence from a large randomized trial. *Archives of general psychiatry*, *65*(5), 578-584. doi:10.1001/archpsyc.65.5.578

Kramer, M. S., Fombonne, E., Igumnov, S., Vanilovich, I., Matush, L., Mironova, E., ... & Promotion of Breastfeeding Intervention Trial (PROBIT) Study Group. (2008b). Effects of prolonged and exclusive breastfeeding on child behavior and maternal adjustment: evidence from a large, randomized trial. *Pediatrics*, *121*(3), e435-e440.

Kurz, C. F. (2022). Augmented inverse probability weighting and the double robustness property. *Medical Decision Making*, *42*(2), 156-167.

Lee, B. K., Lessler, J., & Stuart, E. A. (2010). Improving propensity score weighting using machine learning. *Statistics in medicine*, *29*(3), 337-346.

Li, R., C. Ogden, C. Ballew, C. Gillespie, and L. Grummer-Strawn. 2002. Prevalence of Exclusive Breastfeeding among US Infants: The Third National Health and Nutrition Examination Survey (phase II, 1991–1994). *American Journal of Public Health*, *92*(7), 1107–1110.

MacKinnon, J. G. (2006). Bootstrap methods in econometrics. *Economic Record*, *82*, S2-S18.

Mainieri, T. (2006). The panel study of income dynamics child development supplement: User guide for CDS-II. *Ann Arbor, MI: Institute for Social Research, University of Michigan*, *4*.

Maluccio, J. A., Hoddinott, J., Behrman, J. R., Martorell, R., Quisumbing, A. R., & Stein, A. D. (2009). The impact of improving nutrition during early childhood on education among Guatemalan adults. *The Economic Journal*, *119*(537), 734-763.

Martin, C. R., Ling, P., and Blackburn, G. L. (2016). Review of infant feeding: key features of breast milk and infant formula. *Nutrients*, 8(5), 279.

McCaffrey, D. F., Griffin, B. A., Almirall, D., Slaughter, M. E., Ramchand, R., & Burgette, L. F. (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine*, *32*(19), 3388-3414.

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. Psychological Methods, 9(4), 403–425. https://doi.org/10.1037/1082-989X.9.4.403

McCrory, C., Layte, R., 2011. The effect of breastfeeding on children's educational test scores at nine years of age: results of an Irish cohort study. *Social Science & Medicine*, *72*, 1515–1521.

McGonagle, K. A., & Sastry, N. (2015). Cohort profile: the panel study of income dynamics' child development supplement and transition into adulthood study. *International journal of epidemiology*, *44*(2), 415-422.

McGonagle, K., Schoeni, R., Sastry, N., & Freedman, V. (2012). The Panel Study of Income Dynamics: Child Development Supplement User Guide for CDS-III. *Institute for Social Research: Ann Arbor, MI, USA*.

Meek, J. Y., Noble, L., & Section on Breastfeeding. (2022). Policy statement: breastfeeding and the use of human milk. *Pediatrics*, *150*(1), e2022057988.

Metzger, M. W., & McDade, T. W. (2010). Breastfeeding as obesity prevention in the United States: a sibling difference model. *American Journal of Human Biology*, *22*(3), 291-296.

Moodie, E. E., Saarela, O., & Stephens, D. A. (2018). A doubly robust weighting estimator of the average treatment effect on the treated. *Stat*, *7*(1), e205.

Mohammed, S., Webb, E. L., Calvert, C., Glynn, J. R., Sunny, B. S., Crampin, A. C., … & Oakley, L., L. (2023). Effects of exclusive breastfeeding on educational attainment and longitudinal trajectories of grade progression among children in a 13-year follow-up study in Malawi. *Scientific reports*, 13(1), 11413.

Mullainathan, S., & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87-106.

National Center for Health Statistics. 2002. “2000 CDC Growth Charts for the United States: Methods and Development.” Vital and Health Statistics, ser. 11, no. 246. Hyattsville, MD: Centers for Disease Control and Prevention. http://www.cdc.gov/growthcharts/2000growthchart-us.pdf

Peterson, J. L., & Zill, N. (1986). Marital disruption, parent-child relationships, and behavior problems in children. *Journal of Marriage and the Family*, *48*(2), 295-307.

Raissian, K. M., & Su, J. H. (2018). The best of intentions: Prenatal breastfeeding intentions and infant health. *SSM Population Health*, 5, 86-100.

Rees, D. I., & Sabia, J. J. (2009). The effect of breast feeding on educational attainment: Evidence from sibling data. *Journal of Human Capital*, *3*(1), 43-72.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association Section on Bayesian Statistical Science* 1999, pp. 6–10

Robins, J. M., & Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, *90*(429), 122-129.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the american statistical association*, *90*(429), 106-121.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55.

Rothstein, D. S. (2013). Breastfeeding and children's early cognitive outcomes. *Review of Economics and Statistics*, *95*(3), 919-931.

Sacker, A., Quigley, M. A., & Kelly, Y. J. (2006). Breastfeeding and developmental delay: findings from the millennium cohort study. *Pediatrics*, *118*(3), e682-e689.

Scharfstein, D. O., Rotnitzky, A., & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, *94*(448), 1096-1120.

Schultz, T. P. (2002). Wage gains associated with height as a form of health human capital. *American Economic Review*, *92*(2), 349-353.

Soled, D., Keim, S. A., Rapoport, E., Rosen, L., & Adesman, A. (2021). Breastfeeding is associated with a reduced risk of attention-deficit/hyperactivity disorder among preschool children. *Journal of Developmental & Behavioral Pediatrics*, *42*(1), 9-15.

Snetselaar, L. G., de Jesus, J. M., DeSilva, D. M., & Stoody, E. E. (2021). Dietary guidelines for Americans, 2020-2025: understanding the scientific process, guidelines, and key recommendations. *Nutrition today*, 56(6), 287-295.

US Department of Health and Human Services. (2011). The Surgeon General's call to action to support breastfeeding 2011.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of economic perspectives*, *28*(2), 3-28.

Victora, C. G., Horta, B. L., De Mola, C. L., Quevedo, L., Pinheiro, R. T. Gigante, D. P., … & Barros, R. C. (2015). Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. *The lancet global health*, 3(4), e199-e205.

Victora, C. G., Bahl, R., Barros, A. J., França, G. V., Horton, S., Krasevec, J., ... & Rollins, N. C. (2016). Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect. *The lancet*, *387*(10017), 475-490.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from Head Start. *American Economic Journal: Applied Economics*, *7*(4), 76-102.

Wehby, G. L. (2014). Breastfeeding and child disability: A comparison of siblings from the United States. *Economics & Human Biology*, *15*, 13-22.

Woodcock, R. W. (1997). The Woodcock-Johnson Tests of Cognitive Ability—Revised. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 230–246). The Guilford Press.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel Data, Second Edition. MIT Press.

Wüthrich, K., & Zhu, Y. (2021). Omitted variable bias of Lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics*, 1-47.

Zhao, Z. (2008). Sensitivity of propensity score methods to the specifications. *Economics Letters*, 98(3), 309-319.

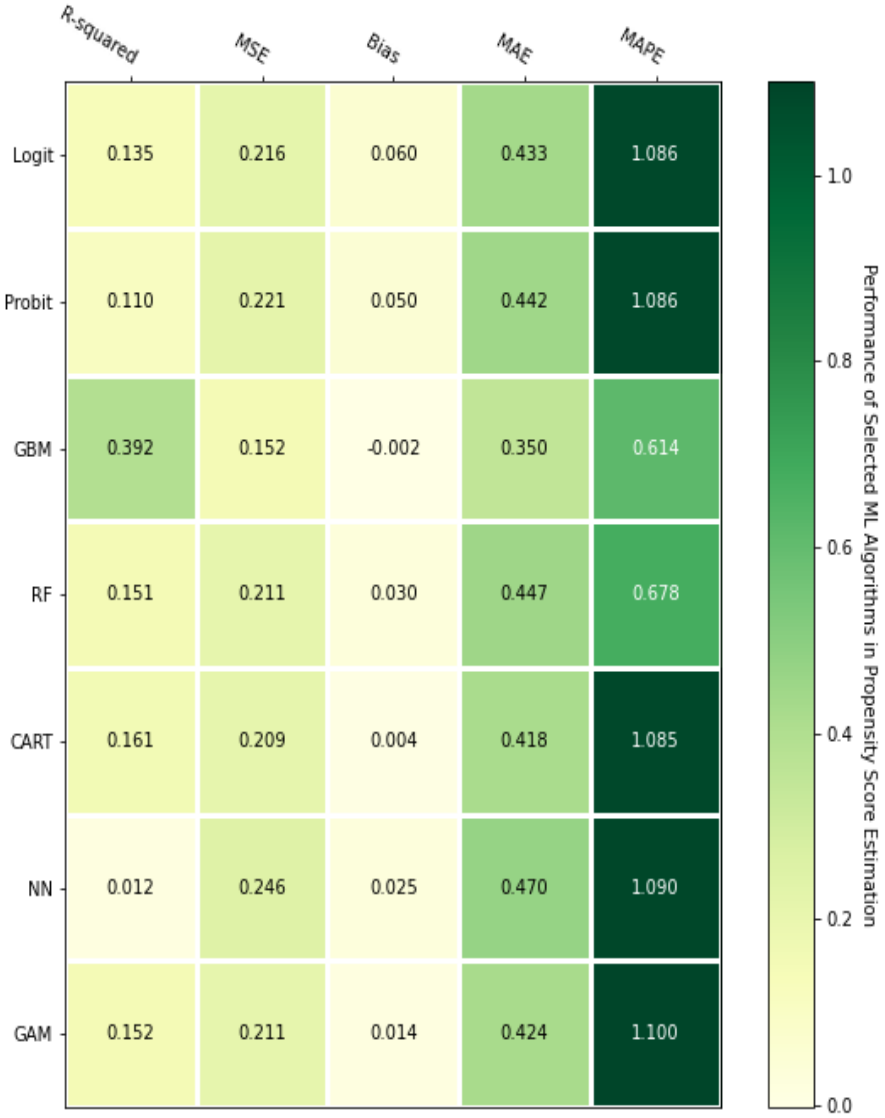


Fig 1. Performance of different machine learning algorithms in propensity score estimation in the DR process

Table 1. Descriptive statistics of outcomes

| | Full sample | | | Not breastfed | | Breastfed | |
|---|---|---|---|---|---|---|---|
| | Obs | Mean | SD | Mean | SD | Mean | SD |
| *Cognitive outcomes* | | | | | | | |
| Letter word | | | | | | | |
| Wave 2 | 2633 | 103.554 | 18.966 | 101.057 | 18.147 | 109.813 | 18.707 |
| Wave 3 | 1505 | 100.706 | 16.61 | 99.402 | 16.327 | 106.823 | 16.108 |
| Applied problem | | | | | | | |
| Wave 2 | 2625 | 101.909 | 16.973 | 100.319 | 16.628 | 108.166 | 16.932 |
| Wave 3 | 1500 | 102.829 | 15.477 | 101.631 | 15.460 | 109.681 | 14.900 |
| Passage comprehension | | | | | | | |
| Wave 2 | 2541 | 103.61 | 15.897 | 101.654 | 15.075 | 108.223 | 15.233 |
| Wave 3 | 1456 | 97.554 | 14.484 | 96.019 | 14.261 | 102.373 | 14.872 |
| Broad reading | | | | | | | |
| Wave 2 | 2637 | 103.652 | 17.55 | 101.194 | 16.875 | 109.806 | 17.042 |
| Wave 3 | 1506 | 99.044 | 16.049 | 97.506 | 15.823 | 105.092 | 15.559 |
| *Non-cognitive outcomes* | | | | | | | |
| Child BMI z-score | | | | | | | |
| Wave 2 | 2599 | 0.589 | 1.200 | 0.611 | 1.198 | 0.438 | 1.169 |
| Wave 3 | 1434 | 1.141 | 1.654 | 1.302 | 1.779 | 0.826 | 1.446 |
| Obesity | | | | | | | |
| Wave 2 | 2607 | 0.205 | 0.404 | 0.215 | 0.411 | 0.159 | 0.367 |
| Wave 3 | 1440 | 0.235 | 0.424 | 0.296 | 0.457 | 0.161 | 0.368 |
| Hyperactivity | | | | | | | |
| Wave 2 | 2906 | 0.073 | 0.260 | 0.093 | 0.290 | 0.070 | 0.256 |
| Wave 3 | 1506 | 0.097 | 0.296 | 0.135 | 0.341 | 0.071 | 0.257 |
| Behavior problems index | | | | | | | |
| Wave 2 | 2907 | 9.814 | 11.790 | 9.925 | 10.761 | 9.776 | 12.462 |
| Wave 3 | 1504 | 8.836 | 11.184 | 10.228 | 14.966 | 8.326 | 10.370 |

*Notes:* All estimates are sample weight adjusted.

Table 2. Descriptive statistics of independent variables.

| | All cases | | Not breastfed | | Breastfed | | MD | *t*-stat |
|---|---|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD | | |
| *Coefficient of interest* | | | | | | | | |
| Ever breastfed | 0.444 | 0.497 | | | | | | |
| *Controls* | | | | | | | | |
| *Child characteristics* | | | | | | | | |
| Female child | 0.490 | 0.500 | 0.495 | 0.500 | 0.506 | 0.500 | 0.011 | 0.660 |
| Child's age | 8.120 | 2.963 | 8.153 | 2.929 | 8.205 | 2.958 | 0.052 | 0.410 |
| First born | 0.400 | 0.477 | 0.340 | 0.474 | 0.358 | 0.480 | 0.019 | 1.140 |
| Second born | 0.350 | 0.477 | 0.341 | 0.474 | 0.348 | 0.477 | 0.007 | 0.430 |
| Birth weight | 6.829 | 1.436 | 6.777 | 1.399 | 7.122 | 1.295 | 0.345 | 7.510*** |
| Born preterm | 0.113 | 0.317 | 0.132 | 0.339 | 0.104 | 0.306 | -0.028 | -0.840 |
| Born SGA[a] | 0.153 | 0.355 | 0.204 | 0.397 | 0.106 | 0.304 | -0.098 | -1.601 |
| Same health[b] | 0.636 | 0.481 | 0.673 | 0.469 | 0.566 | 0.496 | -0.107 | -6.480*** |
| Better health[b] | 0.278 | 0.448 | 0.216 | 0.411 | 0.367 | 0.482 | 0.152 | 9.820*** |
| HOME measure[c] | 18.02 | 3.152 | 17.621 | 3.08 | 18.527 | 3.112 | 0.906 | 7.460*** |
| *Maternal & family characteristics* | | | | | | | | |
| Mother's age[d] | 27.296 | 5.165 | 27.535 | 5.350 | 27.117 | 4.900 | -0.418 | -2.340** |
| IQ | 91.689 | 17.040 | 90.812 | 16.356 | 97.802 | 17.665 | 6.990 | 10.510*** |
| High school dropout | 0.150 | 0.354 | 0.204 | 0.403 | 0.082 | 0.259 | -0.122 | -6.303*** |
| High school | 0.348 | 0.445 | 0.418 | 0.451 | 0.266 | 0.442 | -0.152 | -9.102*** |
| College or more | 0.335 | 0.472 | 0.314 | 0.464 | 0.354 | 0.478 | 0.040 | 2.394*** |
| Employed mother [d] | 0.643 | 0.479 | 0.670 | 0.471 | 0.627 | 0.484 | -0.043 | -2.180** |
| Non-Hispanic black | 0.400 | 0.380 | 0.068 | 0.251 | 0.044 | 0.206 | -0.023 | -2.520** |
| Other Race/ Ethnicity | 0.130 | 0.320 | 0.034 | 0.182 | 0.035 | 0.183 | 0.001 | 0.080 |
| Married [d] | 0.656 | 0.400 | 0.523 | 0.402 | 0.795 | 0.396 | 0.007 | 0.530 |
| Logged family income [d] | 10.595 | 0.941 | 10.560 | 0.868 | 10.630 | 0.945 | 0.070 | 2.212** |
| Bad health[d] | 0.388 | 0.487 | 0.286 | 0.452 | 0.190 | 0.393 | -0.095 | -3.870*** |
| Smoked[e] | 0.446 | 0.498 | 0.614 | 0.492 | 0.418 | 0.502 | -0.197 | -1.690* |
| Alcohol consumption[e] | 0.252 | 0.435 | 0.317 | 0.468 | 0.324 | 0.471 | 0.007 | 0.100 |
| WIC participation[e] | 0.433 | 0.496 | 0.439 | 0.496 | 0.262 | 0.440 | -0.176 | -10.790*** |
| Received food stamps[e] | 0.223 | 0.416 | 0.215 | 0.411 | 0.121 | 0.326 | -0.094 | -7.360*** |
| Received Medicaid[e] | 0.354 | 0.478 | 0.351 | 0.477 | 0.214 | 0.410 | -0.137 | -8.850*** |
| Area of residence[f] | 0.655 | 0.475 | 0.651 | 0.477 | 0.714 | 0.452 | 0.063 | 3.900*** |
| Region[g] | 2.489 | 1.009 | 2.497 | 0.996 | 2.488 | 1.018 | -.009 | -0.270 |

*Notes:* All estimates are sample weight adjusted. MD indicates the differences in means between breastfed and not breastfed. [a] whether the child was born small for gestational age (SGA), e.g., weighing less than a specified percentile of birth weight for a given gestational age following the gender-specific SGA measure from Alexander et al. (1996). [b] As compared to other babies born at birth. [c] The Home Observation for Measurement of the Environment (HOME) scale taken from Caldwell & Bradley (1984) measures the cognitive stimulation and emotional support parents provide to children. [d] year of childbirth. [e] during pregnancy. [f] 0 =non-metropolitan area, 1= metropolitan area. [g] 1= Northeast, 2= North Central, 3= South, 4 = West, 5 =Alaska, Hawaii. State dummies are considered, but not reported here for brevity. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 3. Breastfeeding initiation (ever breastfed) and child development outcomes.

| | OLS | PSM-GBM | DR-GBM |
|---|---|---|---|
| Cognitive outcomes | | | |
| *Letter word* | | | |
| Wave 2 | 4.472*** | 3.545*** | 3.093*** |
| | (0.970) | (1.090) | (0.879) |
| Wave 3 | 3.315*** | 3.020** | 1.603* |
| | (1.095) | (1.120) | (0.973) |
| *Applied problem* | | | |
| Wave 2 | 5.602*** | 5.578*** | 3.566*** |
| | (0.857) | (0.923) | (0.931) |
| Wave 3 | 4.802*** | 4.781*** | 2.962*** |
| | (0.995) | (1.026) | (0.966) |
| *Passage comprehension* | | | |
| Wave 2 | 3.755*** | 3.596*** | 3.200*** |
| | (0.783) | (0.843) | (0.741) |
| Wave 3 | 3.469*** | 3.235*** | 3.073*** |
| | (0.928) | (1.217) | (0.749) |
| *Broad reading* | | | |
| Wave 2 | 4.544*** | 2.978*** | 2.632*** |
| | (0.881) | (1.025) | (0.753) |
| Wave 3 | 3.521*** | 2.319* | 2.211** |
| | (1.026) | (1.198) | (1.021) |
| Non-cognitive outcomes | | | |
| *Child BMI z-score* | | | |
| Wave 2 | -0.144** | -0.142** | -0.099* |
| | (0.066) | (0.069) | (0.056) |
| Wave 3 | -0.355** | -0.294** | -0.254** |
| | (0.122) | (0.149) | (0.124) |
| *Obesity* | | | |
| Wave 2 | -0.062*** | -0.060** | -0.037** |
| | (0.022) | (0.023) | (0.018) |
| Wave 3 | -0.104*** | -0.096*** | -0.085*** |
| | (0.030) | (0.036) | (0.025) |
| *Hyperactivity* | | | |
| Wave 2 | -0.026* | -0.024 | -0.025* |
| | (0.015) | (0.017) | (0.014) |
| Wave 3 | -0.067*** | -0.058** | -0.036** |
| | (0.019) | (0.027) | (0.016) |
| *Behavior problems index* | | | |
| Wave 2 | 0.569 | 0.935 | 0.823 |
| | (0.667) | (0.675) | (0.556) |
| Wave 3 | -0.766 | -0.872 | -0.630 |
| | (0.923) | (0.824) | (0.566) |

*Notes:* Robust standard errors are provided for OLS, while cluster-robust standard errors are reported for the PSM. The DR standard errors are computed by bootstrapping with 999 repetitions. The number of replications with values one less than a multiple of 100 are preferred to avoid interpolations when using the percentiles as confidence interval limits (MacKinnon, 2006). All standard errors are presented in parentheses. For PSM and DR estimations, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. $^{*}$ $p < 0.1$ $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

Table 4. Sibling fixed effects estimates of the effects of breastfeeding initiation on child development outcomes.

| | Wave 2 | Wave 3 |
|---|---|---|
| Cognitive outcomes | | |
| *Letter word* | 3.309*** | 2.207* |
| | (1.031) | (1.161) |
| *Applied problem* | 3.926*** | 3.506*** |
| | (0.957) | (1.097) |
| *Passage comprehension* | 3.263*** | 3.212*** |
| | (0.835) | (0.963) |
| *Broad reading* | 3.022*** | 2.713** |
| | (0.931) | (1.065) |
| Non-cognitive outcomes | | |
| *Child BMI z-score* | -0.142** | -0.265* |
| | (0.068) | (0.137) |
| *Obesity* | -0.043* | -0.094*** |
| | (0.025) | (0.035) |
| *Hyperactivity* | -0.022* | -0.051** |
| | (0.013) | (0.020) |
| *Behavior problems index* | 0.453 | -0.744 |
| | (0.564) | (0.783) |

*Notes:* Robust standard errors are reported for sibling fixed effects and are in parentheses. For each outcome, the covariates from the baseline model are included as controls. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 5. DR- ATT estimates from the heterogeneity analysis by breastfeeding duration

| | ≤3 months | ≤ 6 months | 6+ to 12 months | 12+ months |
|---|---|---|---|---|
| Cognitive outcomes | | | | |
| *Letter word* | | | | |
| Wave 2 | 3.164*** | 4.234** | 5.177*** | 3.032 |
| | (1.181) | (0.928) | (1.358) | (2.437) |
| Wave 3 | 0.887 | 2.201** | 4.400*** | 2.960 |
| | (1.104) | (0.886) | (1.338) | (2.538) |
| *Applied problem* | | | | |
| Wave 2 | 4.023*** | 3.349*** | 5.285*** | 4.720** |
| | (1.040) | (0.857) | (1.352) | (1.903) |
| Wave 3 | 3.546*** | 2.546*** | 4.204*** | 5.427 |
| | (1.038) | (0.821) | (1.603) | (3.403) |
| *Passage comprehension* | | | | |
| Wave 2 | 3.395. *** | 3.292*** | 3.243*** | 3.103* |
| | (0.949) | (0.776) | (1.222) | (1.835) |
| Wave 3 | 2.551** | 3.036*** | 3.085** | 6.281 |
| | (1.044) | (0.774) | (1.430) | (3.837) |
| *Broad reading* | | | | |
| Wave 2 | 3.078*** | 3.244*** | 3.942*** | 3.096 |
| | (1.026) | (0.847) | (1.262) | (2.285) |
| Wave 3 | 1.517 | 2.628*** | 2.649* | 4.838 |
| | (1.054) | (0.846) | (1.355) | (3.204) |
| Non-cognitive outcomes | | | | |
| *Child BMI z-score* | | | | |
| Wave 2 | -0.096 | -0.144** | -0.106 | -0.076 |
| | (0.072) | (0.064) | (0.095) | (0.169) |
| Wave 3 | -0.282** | -0.324*** | -0.253* | -0.089 |
| | (0.119) | (0.095) | (0.146) | (0.361) |
| *Obesity* | | | | |
| Wave 2 | -0.052** | -0.058*** | -0.024 | -0.071 |
| | (0.024) | (0.020) | (0.031) | (0.057) |
| Wave 3 | -0.083*** | -0.096*** | -0.072* | -0.016 |
| | (0.030) | (0.024) | (0.040) | (0.095) |
| *Hyperactivity* | | | | |
| Wave 2 | -0.041*** | -0.028** | -0.020 | -0.018 |
| | (0.013) | (0.014) | (0.021) | (0.030) |
| Wave 3 | -0.055*** | -0.065*** | -0.052* | 0.109 |
| | (0.018) | (0.017) | (0.029) | (0.084) |
| *Behavior problems index* | | | | |
| Wave 2 | 0.462 | 0.604 | 1.368 | 0.531 |
| | (0.711) | (0.609) | (1.061) | (1.783) |
| Wave 3 | -0.464 | -0.756 | -1.304 | 2.817 |
| | (0.881) | (0.655) | (1.195) | (3.671) |

*Notes:* The DR standard errors are computed by bootstrapping with 999 repetitions and are in parentheses. From the DR estimation, ATT values are reported based on the propensity scores estimated using the GBM algorithm. Each breastfeeding duration category is compared with respect to never breastfeeding. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 6. DR-ATT estimates from the heterogeneity analysis by children's age

| | Wave 2 (ages 5-13) | Wave 3 (ages 10-18) |
|---|---|---|
| Cognitive outcomes | | |
| *Letter word* | 3.439*** | 2.547** |
| | (1.144) | (1.071) |
| *Applied problem* | 2.799** | 2.585** |
| | (1.096) | (1.131) |
| *Passage comprehension* | 3.540*** | 3.284** |
| | (1.071) | (1.362) |
| *Broad reading* | 3.651*** | 2.947** |
| | (1.103) | (1.206) |
| Non-cognitive outcomes | | |
| *Child BMI z-score* | -0.094 | -0.277** |
| | (0.083) | (0.108) |
| *Obesity* | -0.037 | -0.061** |
| | (0.028) | (0.029) |
| *Hyperactivity* | -0.079*** | -0.053** |
| | (0.024) | (0.025) |
| *Behavior problems index* | -0.305 | -0.367 |
| | (0.583) | (0.760) |

*Notes:* The DR standard errors are computed by bootstrapping with 999 repetitions and are in parentheses. From the DR estimation, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$

Table 7. DR-ATT estimates from the heterogeneity analysis by maternal race

| | NHW | NHB | *t*-stat (NHW vs NHB) | Others | *t*-stat (NHW vs others) |
|---|---|---|---|---|---|
| Cognitive outcomes | | | | | |
| *Letter word* | | | | | |
| Wave 2 | 3.027*** | 7.430*** | -1.845* | 6.063** | -1.028 |
| | (0.722) | (2.289) | | (2.864) | |
| Wave 3 | 1.425* | 4.585** | -1.595 | 4.856** | -1.605 |
| | (0.754) | (1.837) | | (1.997) | |
| *Applied problem* | | | | | |
| Wave 2 | 3.775*** | 8.043*** | -1.951* | 3.072* | 0.367 |
| | (0.895) | (1.989) | | (1.753) | |
| Wave 3 | 3.269*** | 5.336*** | -1.036 | 2.432** | 0.548 |
| | (1.015) | (1.779) | | (1.159) | |
| *Passage comprehension* | | | | | |
| Wave 2 | 3.154*** | 2.020 | 0.702 | 3.196** | -0.025 |
| | (0.637) | (1.404) | | (1.518) | |
| Wave 3 | 3.121*** | 2.735* | 0.230 | 3.087* | 0.017 |
| | (0.670) | (1.518) | | (1.805) | |
| *Broad reading* | | | | | |
| Wave 2 | 3.258*** | 4.423** | -0.578 | 6.031*** | -1.180 |
| | (0.887) | (1.843) | | (2.169) | |
| Wave 3 | 2.552*** | 3.458** | -0.482 | 5.473*** | -1.361 |
| | (0.690) | (1.730) | | (2.027) | |
| Non-cognitive outcomes | | | | | |
| *Child BMI z-score* | | | | | |
| Wave 2 | -0.066 | -0.027 | -0.255 | -0.245 | 1.062 |
| | (0.055) | (0.144) | | (0.160) | |
| Wave 3 | -0.281*** | -0.114 | -0.884 | -0.326** | 0.259 |
| | (0.083) | (0.168) | | (0.149) | |
| *Obesity* | | | | | |
| Wave 2 | -0.035** | -0.030 | -0.090 | -0.135*** | 2.094** |
| | (0.017) | (0.053) | | (0.045) | |
| Wave 3 | -0.089*** | -0.035 | -1.162 | -0.117** | 0.662 |
| | (0.019) | (0.041) | | (0.040) | |
| *Hyperactivity* | | | | | |
| Wave 2 | -0.039** | -0.058** | 0.664 | -0.002 | -0.844 |
| | (0.016) | (0.023) | | (0.040) | |
| Wave 3 | -0.065*** | -0.078*** | 0.631 | -0.020 | -1.201 |
| | (0.015) | (0.014) | | (0.035) | |
| *Behavior problems index* | | | | | |
| Wave 2 | 0.591 | 0.134 | 0.257 | 0.582 | 0.004 |
| | (0.552) | (1.715) | | (2.336) | |
| Wave 3 | -0.685 | -2.327*** | 1.992** | 0.275 | -0.428 |
| | (0.559) | (0.588) | | (2.095) | |

Notes: The DR standard errors are computed by bootstrapping with 999 repetitions and are in parentheses. From the DR estimation, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. NHW indicates non-Hispanic white, and NHB indicates non-Hispanic black. Reported t-stats are from two-tailed t-tests of differences in ATT estimates across subsamples. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 8. DR-ATT estimates from the heterogeneity analysis by maternal education

| | High school and below | Some college Degree and Higher | *t*-stat |
|---|---|---|---|
| Cognitive outcomes | | | |
| *Letter word* | | | |
| Wave 2 | 2.254** | 4.355*** | 1.511 |
| | (0.940) | (1.010) | |
| Wave 3 | 0.871 | 3.662*** | 1.980** |
| | (0.887) | (1.115) | |
| *Applied problem* | | | |
| Wave 2 | 2.905** | 5.729*** | 1.490 |
| | (1.197) | (1.499) | |
| Wave 3 | 1.986 | 5.647*** | 2.244** |
| | (1.321) | (1.017) | |
| *Passage comprehension* | | | |
| Wave 2 | 1.372* | 5.896*** | 3.877*** |
| | (0.741) | (0.929) | |
| Wave 3 | 0.904 | 5.591*** | 3.050*** |
| | (1.162) | (1.032) | |
| *Broad reading* | | | |
| Wave 2 | 1.691** | 5.071*** | 2.112** |
| | (0.807) | (1.363) | |
| Wave 3 | 1.128 | 4.758*** | 2.681*** |
| | (0.903) | (1.055) | |
| Non-cognitive outcomes | | | |
| *Child BMI z-score* | | | |
| Wave 2 | -0.026 | -0.190*** | 1.630 |
| | (0.072) | (0.072) | |
| Wave 3 | -0.406*** | -0.227** | 1.034 |
| | (0.130) | (0.105) | |
| *Obesity* | | | |
| Wave 2 | -0.024 | -0.037* | 0.379 |
| | (0.025) | (0.022) | |
| Wave 3 | -0.131*** | -0.063** | 1.425 |
| | (0.039) | (0.027) | |
| *Hyperactivity* | | | |
| Wave 2 | -0.022* | -0.036** | 0.652 |
| | (0.013) | (0.016) | |
| Wave 3 | -0.054*** | -0.078*** | 0.840 |
| | (0.019) | (0.021) | |
| *Behavior problems index* | | | |
| Wave 2 | 0.983 | 0.272 | 0.559 |
| | (0.844) | (0.927) | |
| Wave 3 | -0.278 | -0.839 | 0.427 |
| | (0.855) | (0.976) | |

*Notes:* The DR standard errors are computed by bootstrapping with 999 repetitions and are in parentheses. From the DR estimation, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 9. Breastfeeding and child development outcomes excluding low-birth weight and pre-mature babies.

| | Sibling Fixed Effect | PSM-GBM | DR-GBM |
|---|---|---|---|
| Cognitive outcomes | | | |
| *Letter word* | | | |
| Wave 2 | 2.981*** | 4.797*** | 2.880*** |
| | (1.112) | (1.127) | (0.793) |
| Wave 3 | 2.100* | 3.379*** | 0.889 |
| | (1.254) | (1.238) | (0.816) |
| *Applied problem* | | | |
| Wave 2 | 3.800*** | 4.593*** | 3.591*** |
| | (1.013) | (1.277) | (0.746) |
| Wave 3 | 3.490*** | 2.576* | 2.946*** |
| | (1.178) | (1.528) | (0.770) |
| *Passage comprehension* | | | |
| Wave 2 | 3.575*** | 3.954*** | 3.457*** |
| | (1.028) | (0.882) | (0.847) |
| Wave 3 | 2.878** | 3.910*** | 2.804*** |
| | (1.125) | (1.021) | (0.734) |
| *Broad reading* | | | |
| Wave 2 | 2.590** | 4.608*** | 2.113*** |
| | (1.008) | (0.987) | (0.710) |
| Wave 3 | 2.278** | 3.804*** | 1.700* |
| | (1.145) | (1.129) | (0.870) |
| Non-cognitive outcomes | | | |
| *Child BMI z-score* | | | |
| Wave 2 | -0.135* | -0.135* | -0.099* |
| | (0.073) | (0.076) | (0.055) |
| Wave 3 | -0.364** | -0.382** | -0.310** |
| | (0.146) | (0.157) | (0.135) |
| *Obesity* | | | |
| Wave 2 | -0.056** | -0.037 | -0.041** |
| | (0.027) | (0.030) | (0.018) |
| Wave 3 | -0.116*** | -0.101*** | -0.097*** |
| | (0.038) | (0.039) | (0.022) |
| *Hyperactivity* | | | |
| Wave 2 | -0.022 | -0.024 | -0.021 |
| | (0.015) | (0.018) | (0.013) |
| Wave 3 | -0.056** | -0.041* | -0.039** |
| | (0.021) | (0.021) | (0.019) |
| *Behavior problems index* | | | |
| Wave 2 | 0.174 | 0.864 | 0.700 |
| | (0.655) | (0.711) | (0.486) |
| Wave 3 | -1.141 | -0.740 | -0.918 |
| | (0.728) | (0.942) | (0.563) |

*Notes*: Robust standard errors are provided for sibling fixed effects, while cluster-robust standard errors are reported for the PSM. The DR standard errors are computed by bootstrapping with 999 repetitions. All standard errors are presented in parentheses. For PSM and DR estimations, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table 10. Falsification exercise- breastfeeding and placebo outcomes

| | Sibling Fixed Effect | PSM-GBM | DR-GBM |
|---|---|---|---|
| Birth weight | 0.402 | 0.179 | 0.130 |
| | (0.246) | (0.265) | (0.236) |
| Smoked previously | -0.004 | 0.028 | -0.003 |
| | (0.058) | (0.045) | (0.038) |
| Smoked during pregnancy | 0.030 | -0.006 | -0.041 |
| | (0.161) | (0.004) | (0.142) |
| Drank alcohol previously | -0.044 | -0.060 | -0.055 |
| | (0.080) | (0.092) | (0.057) |
| Drank alcohol during pregnancy | -0.061 | 0.032 | 0.047 |
| | (0.136) | (0.176) | (0.080) |

*Notes:* Robust standard errors are provided for sibling fixed effects, while cluster-robust standard errors are reported for the PSM. The DR standard errors are computed by bootstrapping with 999 repetitions. All standard errors are presented in parentheses. For PSM and DR estimations, ATT values are reported based on the propensity scores that are estimated using the GBM algorithm. $^{*}$ $p < 0.1$ $^{**}$ $p < 0.05$, $^{***}$ $p < 0.01$.

## Appendix A

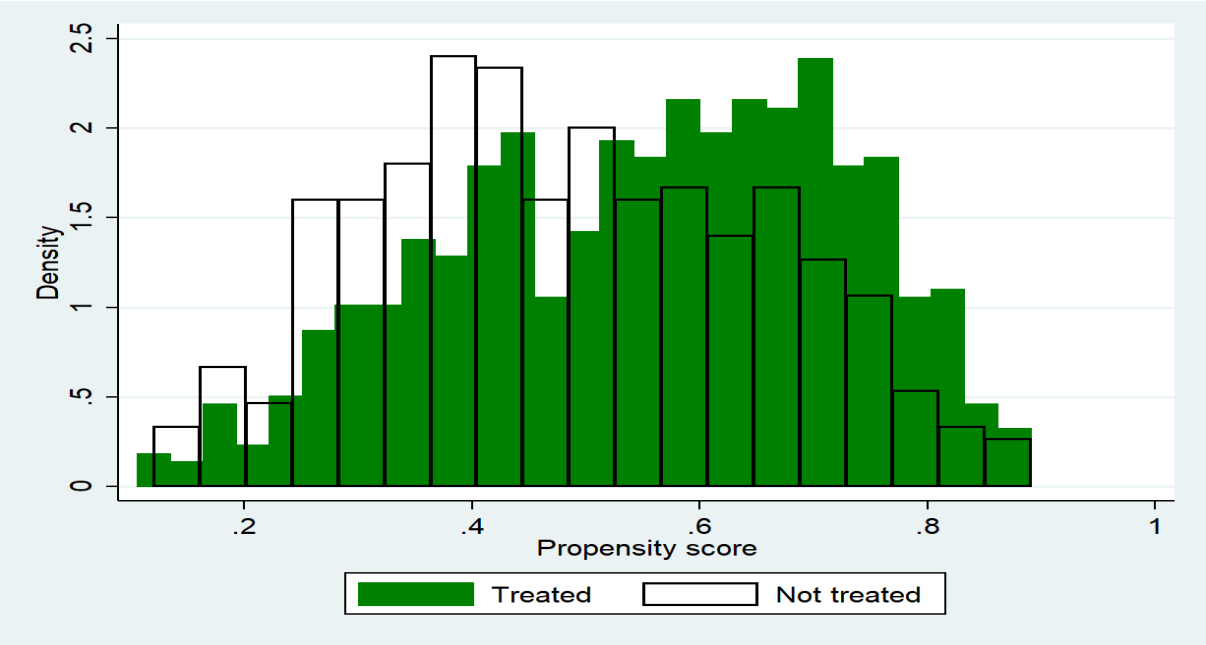


Figure A1. Common support and distributions of propensity score estimates by treatment status.

Table A1. Quality of matching procedures

| Cognitive outcomes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Wave 2 | | | | Wave 3 | | | |
| | BF | NBF | % Bias | *t-stat* | BF | NBF | % Bias | *t-stat* |
| *Child characteristics* | | | | | | | | |
| Female child | 0.494 | 0.476 | 3.400 | 0.520 | 0.563 | 0.491 | 4.400 | 1.520 |
| First born | 0.371 | 0.343 | 5.900 | 0.890 | 0.356 | 0.347 | 1.900 | 0.200 |
| Second born | 0.376 | 0.348 | 5.700 | 0.890 | 0.365 | 0.356 | 1.900 | 0.200 |
| Age | 7.562 | 7.715 | -5.400 | -0.790 | 6.995 | 6.636 | 2.900 | 0.720 |
| Age squared | 65.439 | 68.590 | -6.900 | -1.020 | 55.561 | 50.769 | 1.900 | 0.620 |
| Birth weight | 7.140 | 7.133 | 0.400 | 0.060 | 6.982 | 6.960 | 1.500 | 0.180 |
| Born preterm | 0.083 | 0.107 | -8.300 | -0.410 | 0.073 | 0.091 | -6.400 | -0.340 |
| Born SGA | 0.110 | 0.137 | -6.400 | -0.340 | 0.125 | 0.219 | -3.400 | -0.990 |
| Same health | 0.584 | 0.590 | -1.400 | -0.200 | 0.632 | 0.664 | -6.700 | -0.700 |
| Better health | 0.315 | 0.268 | 1.100 | 1.590 | 0.264 | 0.263 | 0.000 | 0.000 |
| HOME measure | 18.677 | 17.634 | 4.600 | 0.880 | 17.894 | 18.291 | -2.200 | -0.270 |
| *Maternal and family characteristics* | | | | | | | | |
| Age at giving birth | 27.436 | 27.545 | -2.200 | -0.340 | 27.656 | 27.960 | -6.000 | -0.680 |
| Age squared | 776.490 | 782.910 | -2.200 | -0.350 | 788.270 | 803.850 | -5.500 | -0.620 |
| IQ | 102.150 | 101.090 | 0.900 | 0.160 | 105.870 | 104.090 | 1.600 | 0.200 |
| High school dropout | 0.082 | 0.103 | -6.800 | -1.130 | 0.068 | 0.091 | -7.300 | -0.880 |
| High school | 0.373 | 0.401 | -5.700 | -0.870 | 0.400 | 0.414 | -2.800 | -0.290 |
| College or more | 0.545 | 0.496 | 9.900 | 1.510 | 0.518 | 0.509 | -1.800 | 0.190 |
| Employed mother | 0.633 | 0.675 | -8.700 | -1.490 | 0.627 | 0.605 | 4.700 | 0.490 |
| Non-Hispanic Black | 0.023 | 0.032 | -4.600 | -0.910 | 0.018 | 0.014 | 2.300 | 0.380 |
| Other Race/ Ethnicity | 0.030 | 0.025 | 3.100 | 0.550 | 0.027 | 0.045 | -9.400 | -1.020 |
| Married | 0.818 | 0.800 | 4.400 | 0.240 | 0.877 | 0.868 | 2.800 | 0.290 |
| Logged family income | 10.895 | 10.836 | 6.400 | 1.090 | 10.981 | 10.877 | 3.800 | 1.430 |
| Bad health | 0.201 | 0.242 | -3.300 | -1.910* | 0.105 | 0.073 | 1.400 | 1.170 |
| Smoked | 0.433 | 0.600 | -1.800 | -0.710 | 0.417 | 0.416 | 0.000 | 0.000 |
| Alcohol consumption | 0.290 | 0.258 | 7.300 | 0.280 | 0.298 | 0.297 | 0.000 | 0.000 |
| WIC participation | 0.283 | 0.290 | -1.300 | -0.220 | 0.318 | 0.332 | -2.8 | -0.300 |
| Received food stamps | 0.144 | 0.157 | -3.200 | -0.550 | 0.170 | 0.184 | -3.1 | -0.370 |
| Received Medicaid | 0.223 | 0.234 | -2.300 | -0.390 | 0.260 | 0.265 | -1.00 | 0.110 |
| Area of residence | 0.693 | 0.697 | -0.900 | -0.140 | 0.700 | 0.709 | -1.900 | -0.210 |
| Region | 2.469 | 2.507 | -3.800 | -0.640 | 2.444 | 2.372 | 7.2 | 0.750 |
| Non-cognitive outcomes | | | | | | | | |
| | Wave 2 | | | | Wave 3 | | | |
| | BF | NBF | % Bias | *t-stat* | BF | NBF | % Bias | *t-stat* |
| *Child characteristics* | | | | | | | | |
| Female child | 0.490 | 0.442 | 9.600 | 1.460 | 0.493 | 0.507 | -2.900 | -0.290 |
| First born | 0.344 | 0.341 | 0.500 | 0.070 | 0.359 | 0.335 | 5.000 | 0.510 |
| Second born | 0.376 | 0.337 | 8.100 | 1.240 | 0.364 | 0.407 | -8.900 | -0.900 |
| Age | 7.535 | 7.418 | 4.200 | 0.620 | 5.278 | 5.264 | 1.100 | 0.110 |
| Age squared | 65.082 | 63.138 | 4.300 | 0.640 | 29.691 | 29.511 | 1.200 | 0.120 |
| Birth weight | 7.155 | 7.124 | 2.100 | 0.310 | 7.005 | 6.990 | 1.000 | 0.110 |
| Born preterm | 0.107 | 0.137 | -4.500 | -0.520 | 0.129 | 0.226 | -4.200 | -0.990 |
| Born SGA | 0.091 | 0.118 | 5.500 | -0.690 | 0.179 | 0.180 | -0.000 | -0.010 |
| Same health | 0.582 | 0.567 | 3.2 | 0.470 | 0.632 | 0.627 | 1.000 | 0.100 |
| Better health | 0.328 | 0.319 | 2.000 | 0.280 | 0.268 | 0.258 | 2.200 | 0.220 |
| HOME measure | 18.680 | 19.304 | -2.700 | -0.470 | 17.778 | 21.074 | -1.300 | -1.040 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *Maternal and family characteristics* | | | | | | | | |
| Age at giving birth | 27.549 | 27.656 | -2.100 | -0.330 | 27.565 | 27.713 | -3.000 | -0.320 |
| Age squared | 783.150 | 789.670 | -2.300 | -0.350 | 782.860 | 791.030 | -2.900 | -0.310 |
| IQ | 100.660 | 98.720 | 1.600 | 0.270 | 104.850 | 102.130 | 2.400 | 0.280 |
| High school dropout | 0.107 | 0.120 | -4.100 | -0.620 | 0.096 | 0.077 | 6.100 | 0.700 |
| High school | 0.398 | 0.403 | -0.900 | -0.130 | 0.397 | 0.421 | -4.900 | -0.500 |
| College or more | 0.495 | 0.477 | 3.500 | 0.530 | 0.507 | 0.502 | 1.000 | 0.100 |
| Employed mother | 0.632 | 0.643 | -2.300 | -0.340 | 0.632 | 0.636 | -1.000 | -0.100 |
| Non-Hispanic Black | 0.026 | 0.046 | -9.900 | -1.600 | 0.019 | 0.010 | 4.800 | 0.820 |
| Other Race/ Ethnicity | 0.028 | 0.030 | -1.200 | -0.200 | 0.024 | 0.033 | -5.100 | -0.580 |
| Married | 0.876 | 0.874 | 0.700 | 0.100 | 0.876 | 0.866 | 2.900 | 0.290 |
| Logged family income | 10.947 | 10.946 | 0.100 | 0.020 | 10.966 | 10.907 | 8.199 | 0.860 |
| Bad health | 0.098 | 0.105 | -2.200 | -0.330 | 0.101 | 0.115 | -4.700 | -0.470 |
| Smoked | 0.500 | 0.375 | 23.800 | 0.500 | 0.400 | 0.275 | 13.800 | 0.500 |
| Alcohol consumption | 0.290 | 0.240 | 11.300 | 0.420 | 0.390 | 0.140 | 11.300 | 0.420 |
| WIC participation | 0.287 | 0.289 | -0.500 | -0.070 | 0.325 | 0.340 | -3.000 | -0.310 |
| Received food stamps | 0.140 | 0.133 | 1.6 | 0.290 | 0.177 | 0.148 | 6.700 | 0.790 |
| Received Medicaid | 0.234 | 0.228 | 1.4 | 0.240 | 0.268 | 0.230 | 8.100 | 0.900 |
| Area of residence | 0.691 | 0.652 | 8.4 | 1.270 | 0.702 | 0.697 | 1.000 | 0.110 |
| Region | 2.488 | 2.580 | -9.000 | -1.370 | 2.414 | 2.423 | -1.000 | -0.100 |

Notes: The variable means correspond to samples resulting from nearest-neighbor matching. The different outcome variables have somewhat different numbers of non-missing values. Means reported here correspond to samples with non-missing values of letter word test (as representative of the cognitive outcomes) and child BMI Z-score (as representative of the non-cognitive outcomes). Covariate balance for other outcomes is comparable and available upon request. $^{*}$ $p < 0.10$.

Table A2. Specifications of algorithms used in the PS estimation for the DR technique.

| **Algorithm** | **Specifications** |
|---|---|
| Logistic / Probit regression | The number of iterations: 4 |
| CART | Regularization / complexity parameter ($\alpha$): 0.007<br>Maximum depth: 3<br>Number of splits: 5<br>Terminal nodes: 6<br>Information measure: Gini index |
| Random forest | The number of bootstrapped trees: 500<br>The number of variables attempted for splitting a tree: 16.<br>Information measure: Gini index |
| Gradient boosting | The number of boosted trees: 10,000<br>Maximum depth: 3<br>Learning rate($\eta$): 0.01 |
| Neural network | Size of the structure: 29-13-1<br>● The number of controls: 29<br>● The nodes inside the hidden layer: 13<br>● Outcome: 1<br>Activation function: Sigmoid<br>The value of decay: 0.018 |
| General additive model | Convergence threshold: $10^{-7}$<br>Maximum iterations: 200 |

Notes: All these specifications are based on 10-fold cross validation.

Table A3. Specifications of algorithms used in the PS estimation for the PSM technique.

| **Algorithm** | **Specifications** |
|---|---|
| Logistic / Probit regression | The number of iterations: 4 |
| CART | Regularization / complexity parameter ($\alpha$): 0.01<br>Maximum depth: 3<br>Number of splits: 7<br>Terminal nodes: 8<br>Information measure: Gini index |
| Random forest | The number of bootstrapped trees: 500<br>The number of variables tried for splitting a tree: 4<br>Information measure: Gini index |
| Gradient boosting | The number of boosted trees: 10,000<br>Maximum depth: 3<br>Learning rate($\boldsymbol{\eta}$): 0.01 |
| Neural network | Size of the structure: 29-13-1<br>● The number of controls: 29<br>● The nodes inside the hidden layer: 13<br>● Outcome: 1<br>Activation function: Sigmoid<br>The value of decay: 0.01 |
| General additive model | Convergence threshold: $10^{-7}$<br>Maximum iterations: 200 |

Notes: All these specifications are based on 10-fold cross validation.

Table A4. DR-ATT estimates using traditional models and ML algorithms.

| | Logit | Probit | GBM | RF | CART | NN | GAM |
|---|---|---|---|---|---|---|---|
| Cognitive outcomes | | | | | | | |
| *Letter word* | | | | | | | |
| 2002 | 4.229*** | 4.348*** | 3.093*** | 3.295*** | 3.274*** | 3.213*** | 4.228*** |
| | (1.293) | (1.293) | (0.879) | (0.989) | (1.145) | (1.145) | (1.259) |
| 2007 | 2.713** | 2.727** | 1.603* | 2.701*** | 2.616** | 2.923** | 2.715** |
| | (1.212) | (1.212) | (0.973) | (0.985) | (1.145) | (1.178) | (1.156) |
| *Applied problem score* | | | | | | | |
| 2002 | 4.689*** | 4.764*** | 3.566*** | 3.972*** | 4.120*** | 3.848*** | 4.689*** |
| | (1.297) | (1.297) | (0.931) | (0.902) | (1.023) | (1.158) | (1.176) |
| 2007 | 3.260** | 3.322** | 2.794*** | 3.017*** | 2.974** | 3.963** | 3.269** |
| | (1.340) | (1.339) | (0.922) | (0.911) | (1.131) | (1.282) | (1.372) |
| *Passage comprehension* | | | | | | | |
| 2002 | 4.326*** | 4.345*** | 3.206*** | 3.573*** | 4.156*** | 3.994*** | 4.326*** |
| | (0.862) | (0.862) | (0.676) | (0.762) | (0.950) | (0.836) | (0.905) |
| 2007 | 3.114*** | 3.165*** | 3.049*** | 3.038*** | 2.739** | 2.879** | 3.114** |
| | (1.156) | (1.147) | (0.809) | (0.930) | (1.161) | (1.251) | (1.217) |
| *Broad reading* | | | | | | | |
| 2002 | 3.029*** | 3.049*** | 2.532*** | 2.788*** | 2.902*** | 3.324*** | 3.028*** |
| | (1.004) | (1.004) | (0.767) | (0.856) | (1.086) | (1.114) | (1.011) |
| 2007 | 2.247* | 2.312* | 2.129** | 2.313** | 2.045* | 1.509 | 2.223* |
| | (1.295) | (1.295) | (0.869) | (0.955) | (1.185) | (1.481) | (1.262) |
| Non-cognitive outcomes | | | | | | | |
| *Child BMI z-score* | | | | | | | |
| 2002 | -0.118* | -0.118* | -0.099* | -0.129** | -0.140* | -0.138* | -0.117 |
| | (0.072) | (0.071) | (0.056) | (0.062) | (0.077) | (0.072) | (0.072) |
| 2007 | -0.338*** | -0.338*** | -0.254** | -0.337*** | -0.324** | -0.374*** | -0.338*** |
| | (0.130) | (0.133) | (0.124) | (0.108) | (0.127) | (0.134) | (0.131) |
| *Obesity* | | | | | | | |
| 2002 | -0.046* | -0.046** | -0.037** | -0.051** | -0.056** | -0.067** | -0.046** |
| | (0.022) | (0.022) | (0.018) | (0.020) | (0.025) | (0.027) | (0.023) |
| 2007 | -0.089*** | -0.088*** | -0.085*** | -0.094*** | -0.092*** | -0.095*** | -0.089*** |
| | (0.031) | (0.031) | (0.025) | (0.025) | (0.033) | (0.028) | (0.031) |
| *Hyperactivity* | | | | | | | |
| 2002 | -0.026* | -0.027* | -0.025* | -0.022* | -0.019 | -0.022 | -0.020 |
| | (0.014) | (0.014) | (0.014) | (0.013) | (0.016) | (0.014) | (0.014) |
| 2007 | -0.043* | -0.043* | -0.036** | -0.043** | -0.043** | -0.008 | -0.043* |
| | (0.023) | (0.023) | (0.016) | (0.018) | (0.021) | (0.030) | (0.023) |
| *Behavior problem index* | | | | | | | |
| 2002 | 0.884 | 0.878 | 0.823 | 0.993* | 0.983 | 1.036* | 0.884 |
| | (0.654) | (0.654) | (0.556) | (0.582) | (0.762) | (0.631) | (0.630) |
| 2007 | -0.713 | -0.680 | -0.630 | -0.709 | -0.720 | -0.766 | -0.720 |
| | (0.750) | (0.750) | (0.566) | (0.753) | (0.957) | (0.809) | (0.975) |

Notes: All standard errors are in parentheses computed by bootstrapping with 999 repetitions. For each outcome, the covariates from the baseline model are included as controls. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

Table A5. PSM-ATT estimates using traditional models and ML algorithms.

| | Logit | Probit | GBM | RF | CART | NN | GAM |
|---|---|---|---|---|---|---|---|
| Cognitive outcomes | | | | | | | |
| *Letter word* | | | | | | | |
| 2002 | 4.847*** | 4.836*** | 3.545*** | 4.817*** | 4.862*** | 5.796*** | 4.847*** |
| | (0.998) | (1.020) | (1.090) | (1.019) | (0.994) | (1.005) | (0.998) |
| 2007 | 3.172*** | 3.172*** | 3.020*** | 3.180*** | 3.198*** | 3.156*** | 3.172*** |
| | (1.141) | (1.115) | (1.122) | (1.140) | (1.081) | (1.125) | (1.141) |
| *Applied problem score* | | | | | | | |
| 2002 | 6.079*** | 6.098*** | 4.287*** | 6.072*** | 6.099*** | 6.050*** | 6.079*** |
| | (0.925) | (0.946) | (1.042) | (0.951) | (0.937) | (0.921) | (0.925) |
| 2007 | 4.680*** | 4.680*** | 3.201*** | 4.693*** | 4.675*** | 4.550*** | 4.680*** |
| | (1.080) | (1.074) | (1.235) | (1.003) | (0.983) | (1.018) | (1.080) |
| *Passage comprehension* | | | | | | | |
| 2002 | 4.120*** | 4.146*** | 3.602*** | 4.015*** | 4.291*** | 4.150*** | 4.120*** |
| | (0.778) | (0.829) | (0.831) | (0.852) | (0.846) | (0.826) | (0.779) |
| 2007 | 3.916*** | 3.916*** | 3.095*** | 3.872*** | 3.757*** | 3.676*** | 3.916*** |
| | (0.928) | (0.965) | (1.187) | (0.966) | (0.893) | (0.946) | (0.928) |
| *Broad reading* | | | | | | | |
| 2002 | 4.890*** | 4.850*** | 2.915*** | 4.767*** | 5.001*** | 5.050*** | 4.890*** |
| | (0.906) | (0.904) | (1.051) | (0.917) | (0.931) | (0.910) | (0.906) |
| 2007 | 3.641*** | 3.642*** | 2.319* | 3.633*** | 3.581*** | 3.526*** | 3.642*** |
| | (1.040) | (1.040) | (1.198) | (1.076) | (0.997) | (1.051) | (1.040) |
| Non-cognitive outcomes | | | | | | | |
| *Child BMI z-score* | | | | | | | |
| 2002 | -0.152** | -0.155** | -0.142** | -0.134* | -0.120* | -0.141* | -0.152** |
| | (0.070) | (0.073) | (0.069) | (0.073) | (0.069) | (0.071) | (0.070) |
| 2007 | -0.381*** | -0.383*** | -0.294** | -0.378*** | -0.379*** | -0.391*** | -0.381*** |
| | (0.142) | (0.135) | (0.149) | (0.129) | (0.132) | (0.132) | (0.142) |
| *Obesity* | | | | | | | |
| 2002 | -0.061*** | -0.061*** | -0.060** | -0.056** | -0.054** | -0.062*** | -0.061*** |
| | (0.023) | (0.023) | (0.023) | (0.023) | (0.022) | (0.023) | (0.023) |
| 2007 | -0.107*** | -0.107*** | -0.096*** | -0.100*** | -0.101*** | -0.099*** | -0.107*** |
| | (0.034) | (0.033) | (0.036) | (0.032) | (0.032) | (0.033) | (0.034) |
| *Hyperactivity* | | | | | | | |
| 2002 | -0.021 | -0.020 | -0.021 | -0.014 | -0.024* | -0.021 | -0.021 |
| | (0.013) | (0.014) | (0.017) | (0.014) | (0.014) | (0.014) | (0.013) |
| 2007 | -0.044** | -0.044** | -0.042** | -0.041* | -0.042* | -0.042* | -0.044** |
| | (0.022) | (0.021) | (0.021) | (0.021) | (0.022) | (0.022) | (0.022) |
| *Behavior problem index* | | | | | | | |
| 2002 | 0.613 | 0.629 | 0.935 | 0.773 | 0.632 | 0.518 | 0.613 |
| | (0.566) | (0.569) | (0.675) | (0.564) | (0.586) | (0.569) | (0.566) |
| 2007 | -0.994 | -0.977 | -0.872 | -0.719 | -0.832 | -0.751 | -0.944 |
| | (0.748) | (0.748) | (0.824) | (0.757) | (0.739) | (0.730) | (0.748) |

Notes: DR standard errors are in parentheses computed by bootstrapping with 999 repetitions. For each outcome, the covariates from the baseline model are included as controls. * $p < 0.1$ ** $p < 0.05$, *** $p < 0.01$.

## Appendix B

The steps involved in the GBM are below (Friedman, 2001; Hastie et al., 2009; Boehmke & Greenwell, 2019)

Step 1 : Set the problem as $F_0(x) = argmin_\gamma \sum_{i=1}^{N} L(y_i, \gamma)$, where $F_0(x)$ approximates the output and γ represents the learning rate. *L* denotes a loss function, and the mean squared error is commonly used for it. The overall task is to minimize the value of the loss function.

Step 2: For $m = 1$ to *M* (the number of boosting iterations),

Step 2-a: Estimate the 'pseudo-residual' $r_{im}$ by solving the following optimization problem (steepest-descent according to Friedman). *i* represents the observations of training data.

$$r_{im} = -\left|\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right|_{F(x)=F_{m-1}(x)} \quad i = 1, \dots, N$$

Step 2-b: Using the pseudo-residual $r_{im}$ as the output, fit a simple regression tree with the training data. Hastie et al. (2009) claims this step should produce terminal regions (observations that belong to the end node of the tree) $R_{jm}$ and $j = 1,2, \dots, J_m$. The final number of trees built for boosting therefore is $J_m$.

Step 2-c: Once the trees are generated, estimate the learning rate $\gamma_{jm}$ which is between 0 and 1 by solving the following optimization problem. The purpose of this step is to find the optimal learning rate.

$$\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

Step 2-d: Update the tree by placing the learning rate. A smaller learning rate indicates that a longer computation time for boosting and vice versa.

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

Step 3: Output $\widehat{f(x)} = f_M(x)$, which is the final sum of all boosted trees from Step 2-d.

In sum, the algorithm of GBM is to pick an estimation of the output and then minimize the error via boosting. This is done in a sequential matter.