Observational Studies 1 (2015) 30-55 Submitted 5/14; Published 2/15

Application of Propensity Scores to a Continuous Exposure: Effect of Lead Exposure in Early Childhood on Reading and Mathematics Scores

Michael R. Elliott **mrelliot@umich.edu**
Department of Biostatistics
University of Michigan
1420 Washington Heights
Ann Arbor, MI 48109 USA and
Survey Methodology Program, Institute for Social Research
University of Michigan
426 Thompson Street
Ann Arbor, MI 48106 USA

Nanhua Zhang **nanhua.zhang@cchmc.org**
Division of Biostatistics & Epidemiology
Cincinnati Children's Hospital Medical Center
3333 Burnet Avenue
Cincinnati, OH 45229-3026 USA

Dylan S. Small **dsmall@wharton.upenn.edu**
Department of Statistics
The Wharton School, University of Pennsylvania
400 Huntsman Hall, 3730 Walnut St.
Philadelphia, PA 19104 USA

Abstract

The estimation of causal effects in observational studies is usually limited by the lack of randomization, which can result in different treatment or exposure groups differing systematically with respect to characteristics that influence outcomes. To remove such systematic differences, methods to "balance" subjects on observed covariates across treatment or exposure levels have been developed over the past three decades. These methods have been primarily developed in settings with binary treatment or exposures. However, in many observational studies, the exposures are continuous instead of being binary or discrete, and are usually considered as doses of treatment. In this manuscript we consider estimating the causal effect of early childhood lead exposure on youth academic achievement, where the exposure variable blood lead concentration can take any values that are greater than or equal to 0, using three balancing methods: propensity score analysis, non-bipartite matching, and Bayesian regression trees. We find some evidence that the standard logistic regression analysis controlling for age and socioeconomic confounders used in previous analyses (Zhang et al. (2013)) overstates the effect of lead exposure on performance on standardized mathematics and reading examinations; however, significant declines remain, including at doses currently below the recommended exposure levels.

Keywords: Causal inference, non-bipartite analysis, Bayesian regression trees (BART).

1. Introduction

We consider the problem of estimating the causal effect of a continuous dose from an observational study in the context of an analysis that related early childhood lead exposure to elementary school standardized test scores (Zhang et al. 2013). While the effects of lead exposure on cognitive development are understood to some degree, little previous work has documented the effect of such exposure on school performance. In addition, determining the dose effect of lead exposure on school performance rather than just whether it has an effect is important for deciding what level of lead exposure in a child is of concern for public policy. Zhang et al. found the odds of scoring less than proficient on reading and mathematics tests approximately doubled for those whose blood lead levels were greater than 10 micrograms versus those whose blood lead levels were less than 1 micrograms per deciliter. Drawing causal inference from these associations is made difficult by the strong relationship between factors known to be a priori associated with test results – race/ethnicity, socio-economic status, and mother education – and lead exposure. Zhang et al.'s analysis adjusted for these potential confounders as linear terms in a logistic regression model, but did not consider in detail issues such as model misspecification of confounder effects or lack of overlap across levels of these potential confounders and lead exposure. Methods to assess these issues when exposures are continuous are less prevalent than when exposures are binary. We provide a detailed case study of the continuous exposure setting in this manuscript.

1.1 Early Childhood Lead Exposure and Standardized Test Scores Among Schoolchildren in Detroit

Early childhood lead exposure can lead to a wide range of problems in children. The most common effects are subclinical impact on the central nervous system (CNS) (Committee on Environmental Health, 2005), leading to cognitive impairment and behavioral problems. While many studies have focused on the effect of early childhood lead exposure on intellectual and behavioral deficits, very few have related early childhood lead exposure to academic performance in school-aged children. We assess the long-term effect of early childhood lead exposure on academic achievement using data from Detroit, MI, a large city in the US with one of the most severe lead exposure and poisoning problems in children. The blood lead testing surveillance data were collected by the Michigan Department of Community Health (MDCH) Childhood Lead Poisoning Prevention and Control Program (CLPPP). This dataset also contains the identifying variables (first and last name, gender, date of birth) and maternal education. The Detroit Department of Health and Wellness Promotion (DHWP) has childhood lead testing data stored dating back to 1988. For the purposes of this study, we restricted the data to venous blood test results on children 0-6 years of age, born between 1990 and 2008; in the case of children with more than one venous test we used the highest value.

The Michigan Educational Assessment Program, commonly known as MEAP, is a standardized test taken by all public school students in the State of Michigan from elementary school to junior high school. The subjects tested include mathematics, reading, writing, science, and social studies. Mathematics and reading tests are administered from grade 3 through grade 8, writing in grade 4 and 7, science in grade 5 and 8, and social studies in grade 6 and 9. There are four levels for MEAP score categories: level 1 = advanced, level

2 = proficient, level 3=partially proficient and level 4 = not proficient. The data available to us is a dichotomous outcome of "less than proficient" (MEAP level of 3 or 4) vs. "proficient" (MEAP level of 1 or 2). The Detroit Public Schools (DPS) maintains a database with records of all MEAP test results since the MEAP tests were first administered during the 1969-70 school year. This dataset includes identifying information such as name and date of birth. Additionally, the database contains demographics (gender, race, native language) and socioeconomic status (free-lunch status, Medicaid status). We use a subset of this dataset, which contains the MEAP scores covering mathematics and reading in grade 3 and 5 during the years of 2008 to 2010. The two datasets were linked to each other by first and last name, gender and date of birth. A total of 15,652 3rd and 5th grade students were successfully matched to their blood lead surveillance data, which contains information of the highest blood lead level recorded before 6 years old.

1.2 Causal Inference with a Continuous Exposure

The goal of this analysis is to compare the academic scores for students who had different early childhood blood lead levels, but were similar in terms of baseline characteristics. Children can be exposed to lead from a variety of sources, but the two most likely sources in Detroit are soil contamination due to residues from leaded gasoline or previous or ongoing industrial processes in nearby manufacturing facilities, and leaded paint used in older housing stock. Although lead paint was banned from US housing construction as of 1978, more than 90% of the city's housing stock was built before 1980. Differences in these sources of exposure, as well as differences in the degree to which young children play, mouth, or eat these contaminated materials, contribute to variability in childhood lead exposure. Figure 1 shows the distribution of early childhood blood levels. The distribution is highly skewed to the right, with a few very large values of the blood lead concentration. In order to be able to credibly assess the effect of two different blood lead levels on the test scores for children with given baseline characteristics, there need to be children with these baseline characteristics at both of the blood lead levels; otherwise, there is a lack of overlap and causal inferences will involve extrapolation. For example, if the children with low blood lead levels all have high socioeconomic status (SES) and the children with high blood lead levels all have low SES, then any attempt to assess the effect of low vs. high blood levels controlling for SES would involve extrapolation. Another example is that if the children with low blood lead levels have both low and high SES but the children with high blood lead levels have only low SES, then the effect of low vs. high blood lead levels for low SES children can potentially be assessed without extrapolation but any attempt to assess the effect of low vs. high blood lead levels for high SES children would involve extrapolation. Standard multiple regression methods do not make it transparent whether such an extrapolation is occurring.

Here we consider three "state of the art" methods which attempt to make transparent whether there is sufficient overlap for causal inferences and if not, what subset of the dose levels and covariates that causal inference can be made on without extrapolation: (i) subclassification on a propensity score for a continuous exposure (Imai and Van Dyk, 2004); (ii) matching on a propensity score for a continuous exposure using a non-bipartite matching algorithm(Lu et al., 2001; Lu et al., 2011) ; (iii) use of the Bayesian additive regression tree (BART) model (Chipman et al. 2010) that provides a highly flexible yet parsimonious

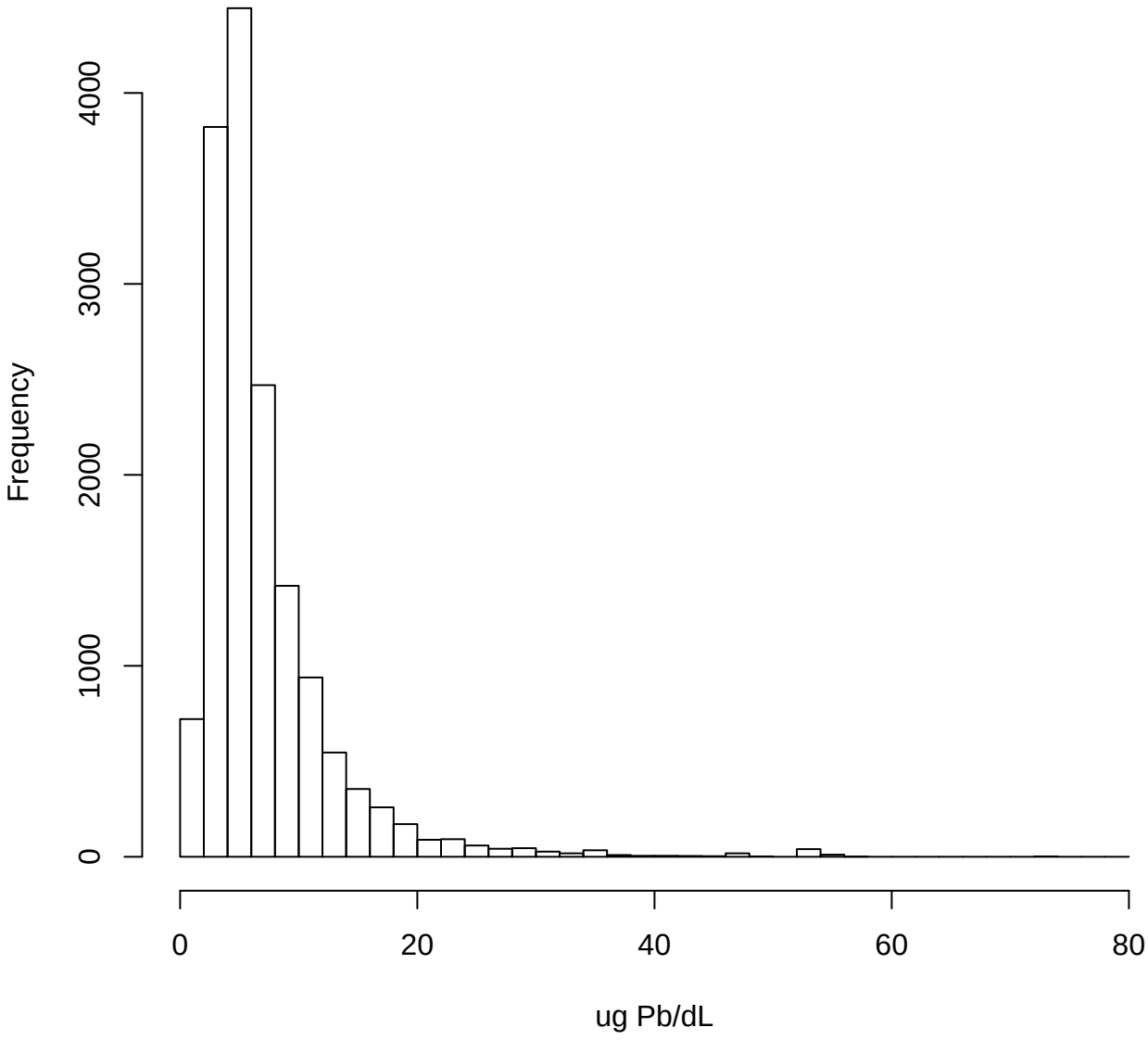

Figure 1: Histogram of lead blood levels.

model to estimate the effect of dose, along with a method for assessing overlap (Hill, 2011 and Hill and Su, 2013). Propensity score methods have been commonly used in the setting of binary exposures, but far less commonly used in the setting of continuous exposures; the BART model has been used by Hill (2011) for causal inference for a binary exposure but has not previously been used for a continuous exposure to our knowledge.

2. Methods for Causal Inference with a Continuous Exposure

2.1 Propensity Score Subclassification with Continuous Exposures

Imai and Van Dyk (2004) develop methodology for extending the propensity score into a setting for continuous treatment. Briefly, propensity score analysis relies on the assumption that, conditional on observed covariates, the "assignment" of the treatment is random. Let t_i denote the observed treatment for the ith observation, where t_i is a realization of a potential assignment $T_i \in \mathcal{T}$, where $\mathcal{T}$ is the sample space of the treatment variable. The set of potential outcomes $\mathcal{Y}_i = \{Y_i(T_i)\}$ is then also of infinite size, though the sample space for the potential outcomes themselves are not necessarily continuous (in our example, the outcomes are binary, so $Y_i(T_i) \in \{0, 1\}$ for $T_i \in \mathcal{T}$). Propensity score analysis obtains causal inference under two key assumptions: (1) stable unit treatment value, i.e., the treatment assignment to the jth subject has no impact on the ith subject ($Y_i(\boldsymbol{T}) = Y_i(T_1, ..., T_n) \equiv Y_i(T_i)$) and the observed outcome for subject i, Y_i, is equal to the potential outcome for subject i under her observed treatment t_i, i.e., $Y_i = Y_i(t_i)$ (Rubin 1980); and (2) strong ignorability of treatment assignment conditional on observed covariates X_i, that is, the observed treatment is independent of the potential outcomes conditional on observed covariates, i.e., $p(T_i = t_i \mid \mathcal{Y}_i, X_i) = p(T_i = t_i \mid X_i)$. Then, for covariates X and treatment levels T^A and T^B such that there is overlap in the sense that $p(T = T^A|X) > 0, p(T = T^B|X) > 0$ where p denotes the probability density function, the average causal effect comparing potential outcomes at treatment T^A with potential outcomes at treatment T^B given by $E(Y_i(T_i^A) - Y_i(T_i^B) \mid X_i)$ can be consistently estimated by the expected values at the observed levels $E(Y_i \mid t_i = T^A, X_i) - E(Y_i \mid t_i = T^B, X_i)$.

Defining the propensity function $e_\theta(X)$ as the conditional probability (density in the case of a continuous treatment) of the treatment given observed covariates parameterized by θ ($e_\theta(X_i) = \{p(T_i \mid X_i; \theta) : T_i \in \mathcal{T}\}$), Imai and Van Dyk show that the propensity function contains all of the information about the distribution of the treatment given X (i.e., it serves as a balancing score: $p(T = t \mid X) = p(T = t \mid e_\theta(X))$), and that, given the propensity function, treatment assignment is strongly ignorable: $p(Y_i(T_i) \mid T_i = t_i, e_\theta(X_i)) = p(Y(T_i) \mid e_\theta(X_i))$ for all $T_i \in \mathcal{T}$. We can then average over $e_\theta(X)$ to obtain $p(Y(T))$ for any specific value of T:

$$p(Y(T)) = \int p(Y(T) \mid e_\theta(X)) p(e_\theta(X)) de_\theta(X) = \int p(Y \mid T, \theta) p(e_\theta(X)) de_\theta(X).$$

This integration can be approximated by classifying observations based on $e_\theta(X)$ into J subsclasses and averaging the distribution of potential outcomes across the subclasses:

$$\int p(Y_i \mid T, \theta) p(e_\theta(X)) de_\theta(X) \approx \sum_{j=1}^{J} p(Y \mid T) W_j \tag{1}$$

where W_j is the proportion of the population in jth subclass. Typically we model $p(Y \mid T) \equiv p(Y \mid T, \phi)$ parametrically, so that the causal effect of interest is ϕ. In practice, we estimate θ from the data and classify accordingly, then compute $\hat{\phi} = \sum_j \hat{\phi}_j \hat{W}_j$, where $\hat{\phi}_j$ is estimated from the model relating Y to t in class j, and $\hat{W}_j$ is based on $\hat{\theta}$.

2.2 Non-bipartite Matching

Under the assumption of 1) strong ignorability, 2) correct modeling of the propensity score $e_\theta(X)$, and 3) an additive relationship between the causal effect of the treatment and the effect of the treatment assignment propensity on the outcome, we have, under a binary outcome for Y,

$$\text{logit}(E(Y_i(T_i) \mid e_\theta(X_i)) = \text{logit}(P(Y_i = 1 \mid T_i, e_\theta(X_i)) = g(e_\theta(X_i), \gamma) + f(T_i, \beta) \tag{2}$$

where $\text{logit}(u) = \log(u/(1-u))$, and $g(\cdot, \gamma)$ and $f(\cdot, \beta)$ are unknown functions of the propensity score and treatment parameterized by γ and β respectively. One option in this setting is to model both $g(e_\theta(X), \gamma)$ and $f(T, \beta)$ fully parametrically; another is to match subjects based on the propensity scores $e_\theta(X)$, forming $j = n/2$ pairs (one subject can be dropped if n is odd). The latter approach replaces the need to correctly estimate g under the assumption that each pair of matched observations has a "base" logit probability b_j of experiencing the event. Conditional logistic regression can then be used to estimate $f(T, \beta)$ by conditioning on $Y_{1j} + Y_{2j}$, where Y_{lj} is the (arbitrarily ordered) lth observation in the jth pair.

When treatment is binary (i.e., a treatment and a control group), then bipartite matching matches treated and control subjects who have similar covariates or propensity scores (Rubin 1973). When there are more than two exposure levels or a continuous exposure, there are no longer two disjoint groups; rather there is a single group and any individual can, in principle, be matched to any other individual (Lu et al., 2001). The goal of the matching when there are more than two exposure levels is to form matched pairs that are similar in terms of covariates but differ markedly in dose. This can be achieved by nonbipartite matching (Lu et al., 2001; Lu et al., 2011).

Non-bipartite matching matches elements without regard to set membership. Optimal matching can be viewed as a minimization problem. Assume we have an even number of nodes $v_1, ..., v_n$ (subjects), and that every potential pairing of nodes $[v_i, v_j]$ $i \neq j$ is a member of the set of edges $\mathcal{E}$, and has a weight w_{ij} (the distance between subjects i and j) associated with it. Denote M as the set of paired nodes $[v_i, v_j]$ between which an edge exists and each node can appear at most once. Optimal matching finds $\boldsymbol{z}$ such that

$$z_{ij} = \begin{cases} 1 & \text{if } [v_i, v_j] \in M \\ 0 & \text{if } [v_i, v_j] \notin M \end{cases}$$

that minimizes

$$D(\boldsymbol{z}) = \sum_{[v_i,v_j]\in\mathcal{E}} w_{ij} z_{ij}$$

subject to $\sum_{j:[v_i,v_j]\in\mathcal{E}} z_{ij} = 1$ for all v_i, $i = 1, ..., n$. Once these matches are made, analyses can proceed by conditional logistic regression as previously noted, treating the pairings as strata.

Both bipartite and non-bipartite matching is optimized by the above criteria; however non-bipartite matching can be difficult to achieve because the size of $\mathcal{E}$ is much larger than in the bipartite setting. We use the R package nbpMatching, described by Lu et al. (2011), that implements an efficient non-bipartite matching algorithm of Derigs (1988). For our distance between subjects i and j (w_{ij}), we construct a scalar propensity score (see Section 3.1) and the distance between i and j is the absolute difference in propensity scores. To increase power by making the pairs far apart in dose, we construct a second matching in which we first divide subjects into subclasses with similar propensity scores and then match subjects within subclasses using as the distance between i and j the absolute difference in i and j's dose.

2.3 Bayesian Additive Regression Trees (BART)

A third method for assessing the causal effect of a continuous treatment is to estimate the counterfactual outcome under a very robust semiparametric model. Hill (2011) proposes the use of Bayesian additive regression trees (BART) (Chipman et al. 1998; Chipman et al. 2010). In classification and regression tree (CART) methodology (Breiman et al. 1984), the distribution of an outcome y given predictors x is determined by partitioning the sample based on x into a series of disjoint nodes in which the values of y are more homogeneous. Various methods which combine a set of tree models, so called ensemble methods, have been developed, e.g., boosting (Freund and Schapire, 1997) and random forests (Breiman, 2001). BART is a Bayesian ensemble method that models the the mean outcome given predictors by a sum of trees; compared to using a single tree to model the outcome, BART can more easily incorporate additive effects of predictors. For a continuous outcome, the basic idea is to approximate the mean of Y given covariates x and treatment T by a sum of m regression trees, $E(Y_i|X_i = x_i, T) \approx \sum_{j=1}^{m} g_j(x_i, T)$ where $g_j(x, T)$ represents the mean assigned to the node in the jth regression tree associated with covariate values x and treatment level T. In our application, where the outcome is binary, the BART approach is to estimate the inverse of the probit function of $E(Y|X, t)$ by a sum of regression tree means:

$$\Phi^{-1}(P(Y_i = 1 \mid t_i, x_i)) = \sum_{j=1}^{m} g(t_i, x_i; R_j, M_j)$$

where R_j denotes the regression tree (actually a set of decision rules defining the nodes of the tree) and M_j is a set of parameter values associated with the terminal nodes of the jth regression tree, $g(t_i, x_i; R_j, M_j)$ maps a parameter value $\mu_{ij} \in M_j$ to each (t_i, x_i) based on the terminal node that (t_i, x_i) falls into in regression tree R_j, and the number of regression trees m is considered fixed and known. Priors $p(M_j \mid R_j)$ and $p(R_j)$ are designed to limit the number of nodes associated with each tree; details can be found in Chipman et

al. Estimation proceeds via a Markov Chain Monte Carlo algorithm, and a draw from the posterior predictive distribution of the mean of the ith outcome $E(Y_i(t_i))$ is obtained via

$$\Phi(\sum_{j=1}^{m} g(t_i, X_i; R_j^{rep}, M_j^{rep})), \tag{3}$$

where (R_j^{rep}, M_j^{rep}) are obtained from the draws from the posterior distribution of (R_j, M_j).

In the causal inference setting, a draw l from the posterior distribution of the expected effect of changing dose from a given level T^A to a different level T^B with covariates x_i is given by

$$d_i^{(l)}(T^A, T^B \mid x_i) = E(Y_i^{(l)}(T^A) \mid x_i) - E(Y_i^{(l)}(T^B) \mid x_i)$$

where the counterfactual $E(Y_i^{(l)}(T) \mid x_i)$ is obtained by replacing (3) with $\Phi(\sum_{j=1}^{m} g(T, X_i; R_j^{(l)}, M_j^{(l)}))$. Using the approximation $E(Y) = \int E(Y \mid X = x)p(X = x)dx \approx \sum_{i=1}^{n} E(Y_i \mid X_i = x_i)$, we then compute a draw from the distribution of the population causal effect on the outcome at dose level T^A versus dose level T^B, $d(T^A, T^B) = E(Y(T^A)) - E(Y(T^B))$, from $d^{(l)}(T^A, T^B) = n^{-1} \sum_{i=1}^{n} d_i^{(l)}(T^A, T^B \mid x_i)$. Estimates of the posterior mean and predictive intervals for $d(T^A, T^B)$ can be obtained based on the empirical mean and order statistics from L draws $d^{(1)}(T^A, T^B), ..., D^{(L)}(T^A, T^B)$. Here we implement BART using the BayesTree package in R.

3. Application to Detroit Blood Lead Level Project

The focus of our analysis is the impact of lead exposure, as measured by blood lead level before age 6, and academic proficiency, as measured by performance on the standardized reading and mathematics exams given annually in Michigan public schools (Michigan Educational Assessment Program [MEAP] tests). We consider the following as potential confounders of this relationship: grade level (3rd vs. 5th), gender, race (black vs. other), English as a second language, family income as measured by eligibility for the free lunch program, and maternal education (more than high school vs. high school or less). Table 1 compares the mean blood lead level (BLL) in μg/dL, along with the proportion achieving proficiency on the MEAP exams, within each level of these potential confounders. As Table 1 shows, there are strong associations between these potential confounders and lead exposure in early childhood, as well as between these potential confounders and reading and mathematics scores, suggesting that these demographic factors may be confounding the relationship between early childhood lead exposure and reading/math scores. Fifth graders had considerably higher levels of blood lead as young children and much lower levels of reading and mathematics proficiency on average than third graders; similarly those who were native English speakers, free-lunch-eligible, and/or whose mothers had a high school education or less were simultaneously more likely to have higher levels of blood lead and lower levels of proficiency at reading and math. Factors such as free-lunch eligibility and maternal education are potentially confounders in that they are socioeconomic status measures that can lower test scores independent of blood lead level, and are associated with factors such as residence in substandard housing that contains lead paint, which can raise blood lead levels.

	Mean log(blood lead level)				% proficient at reading			% proficient at math		
SES measure	Yes	No	Diff(se)	p	Yes	No	p	Yes	No	p
3rd grade (vs. 5th)	1.767	1.882	-.115(.009)	<.001	75.3	62.5	<.001	81.4	52.5	<.001
Male	1.829	1.812	.017(.010)	.069	67.4	71.6	<.001	66.0	70.0	<.001
African-American	1.837	1.678	.159(.016)	<.001	69.5	66.8	.026	67.1	73.5	<.001
ESL	1.776	1.830	-.054(.013)	<.001	71.2	68.9	.021	73.4	66.6	<.001
Free-lunch-Eligible	1.870	1.623	.247(.012)	<.001	67.6	76.2	<.001	66.2	74.0	<.001
Maternal Educ. < HS	1.856	1.704	.153(.011)	<.001	68.0	73.8	<.001	66.7	71.4	<.001

Table 1: Associations between socio-demographic measures and log of blood lead level (μg Pb/dL), and proficient at reading/proficient at mathematics as measured by the Michigan Education Assessment Program (MEAP). ESL=English as a second language. p-value resulting from t-test (blood lead level) or chi-square test (reading and mathematics proficiency).

We first consider three forms of a standard logistic regression analysis to consider the effect of BLL on risk of insufficient proficiency in reading or mathematics: 1) unadjusted, 2) adjusted for all covariates used in the propensity score estimation, and 3) adjusted for covariates used in the propensity score estimation and restricted to those observations with BLLs between 2 and 19 μg/dL – the motivation for restricting to BLLs between 2 and 19 is that these are the values of the BLL for which the probability density of the BLL is not very small for any propensity score subclass, see Section 3.1 below. To allow for a reasonably non-parametric relationship between BLL and risk of insufficient proficiency in reading, we fit a restricted cubic spline (Durrleman and Simon, 1989) with knots $(\kappa_1, ..., \kappa_5)$ at log(BLL+1) of 1.33, 1.67, 2, 2.33, and 2.67. Thus our base model is of the form

$$log\left(\frac{P_i}{1-P_i}\right) = \beta_0 + \beta_1 t_i + \beta_2 S(t_i)_1 + \beta_3 S(t_i)_2 + \beta_4 S(t_i)_3 \tag{4}$$

where P_i is the probability of the ith child having a less than proficient test score and $S(t)_q$ is given by $(t-\kappa_q)_+^3 + \frac{\kappa_5-\kappa_q}{\kappa_5-\kappa_4}(t-\kappa_4)_+^3 + \frac{\kappa_4-\kappa_q}{\kappa_5-\kappa_4}(t-\kappa_5)_+^3$ for $(x)_+ = x$ if $x > 0$ and 0 if $x \leq 0$; we add the term $\gamma' X_i$ to (4) when adjusting for covariates. Table 4 gives the results. Adjusting for potential confounders appears to reduce the impact of BLL on both the reading and mathematics outcomes, although the overall association remains highly significant. Adjusting for confounders and restricting the support to lie between log blood levels of 1 and 3 "compresses" the effect to some degree, although the overall pattern remains.

3.1 Propensity Score Analysis

In our application, we develop a propensity function $p(t \mid X)$ where t is log(BLL+1) by modeling $t \mid X \sim N(X'\alpha, \sigma^2)$. Hence the resulting propensity score is given by $\hat{\theta}(X) = X'\hat{\alpha}$, where $\hat{\alpha}$ is estimated by the linear regression of log(BLL+1) on an intercept as well as covariates X: grade level, gender, race, native English speaker, free-lunch eligibility, and

		Unadjusted	Adjusted	Adjusted Restricted	Propensity	Propensity CL	BLL-maximized Propensity CL
Reading	Intercept	-1.880(.136)	-2.011(.149)	-1.556(.214)	-1.380(.213)	—	—
	Linear	.569(.099)	.478(.100)	.208(.147)	.283(.213)	.414(.150)	.239(.220)
	S1	.558(.380)	.648(.384)	1.714(.544)	1.432(.562)	1.327(.804)	1.573(.892)
	S2	-1.897(1.023)	-2.170(1.034)	-5.247(1.562)	-3.792(1.630)	-3.705(2.315)	-4.982(2.665)
	S3	1.725(.799)	1.963(.808)	4.722(1.432)	3.332(1.538)	2.787(2.128)	4.783(2.526)
Math	Intercept	-1.799(.134)	-2.682(.154)	-2.661(.223)	-1.505(.213)	—	—
	Linear	.542(.098)	.393(.107)	.357(.152)	.352(.151)	.365(.220)	-.040(.237)
	S1	.824(.378)	1.070(.394)	.750(.560)	.665(.578)	.160(.818)	2.381(.955)
	S2	-2.763(1.019)	-3.486(1.067)	-1.973(1.611)	-1.956(1.674)	.274(2.351)	-6.488(2.834)
	S3	2.590(.800)	3.194(.839)	1.078(1.480)	1.436(1.596)	-1.334(2.146)	4.560(2.662)

Table 2: Parameters predicting log odds of insufficient progress in reading/mathematics on Michigan Education Assessment Program as restricted cubic spline function of log blood lead level with knots at 1.33, 1.67, 2, 2.33, and 2.67 log(Pb μg/dL): unadjusted, adjusted for all covariates used in the propensity score estimation, adjusted for covariates used in the propensity score estimation and restricted to those observations with blood lead levels between 2 and 19 μg/dL; along with propensity-score adjusted, propensity-score matched, and propensity-score matched maximized for BLL differences, all constrained to those observations with blood lead levels between 2 and 19 μg/dL. S1, S3, and S3 correspond to the restricted cubic spline bases; see Durrleman and Simon (1989) for details.

maternal education. We considered a main-effects only model as well as models with all two-way, all three-way, through all six-way interactions. Using the AIC criterion, a model with all three-way interactions is chosen as the best fit. Figure 2 shows the distribution of the propensity scores, along with the plot of propensity scores against the log(BLL+1) measures and a residual plot of the difference between the propensity scores and the log(BLL+1). These plots show the discrete nature of the BLL (measured to the nearest μg/dL), but overall a linear link appears a reasonable first approximation to the prediction of log lead level. The degree of variance reduction in the model of log blood lead is highly significant ($F_{41,15610}$=28.82), although the total amount of variance explained is modest (adjusted R^2=.068).

We utilize the propensity score by dividing it into 8 classes based on cutpoints at 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, and 2.1 of log(BLL+1). The distribution of the covariates by the eight propensity score classes is given in Table 2. Although interactions break monotonicity in the prediction of BLL, some strong predictors (grade level, free-lunch eligibility, and maternal education) are very highly associated with propensity score class, with 3rd graders absent from the three highest propensity score levels, and free-lunch eligible students absent from the lowest propensity score level. We consider the degree to which we have successfully balanced the covariates using the propensity score by repeating the t-tests for Table 1 within each of the eight weight classes. The results are shown in Table 3. All associations now fail to reject at $\alpha = .05$, with the exception of grade level in class 3, where the fifth

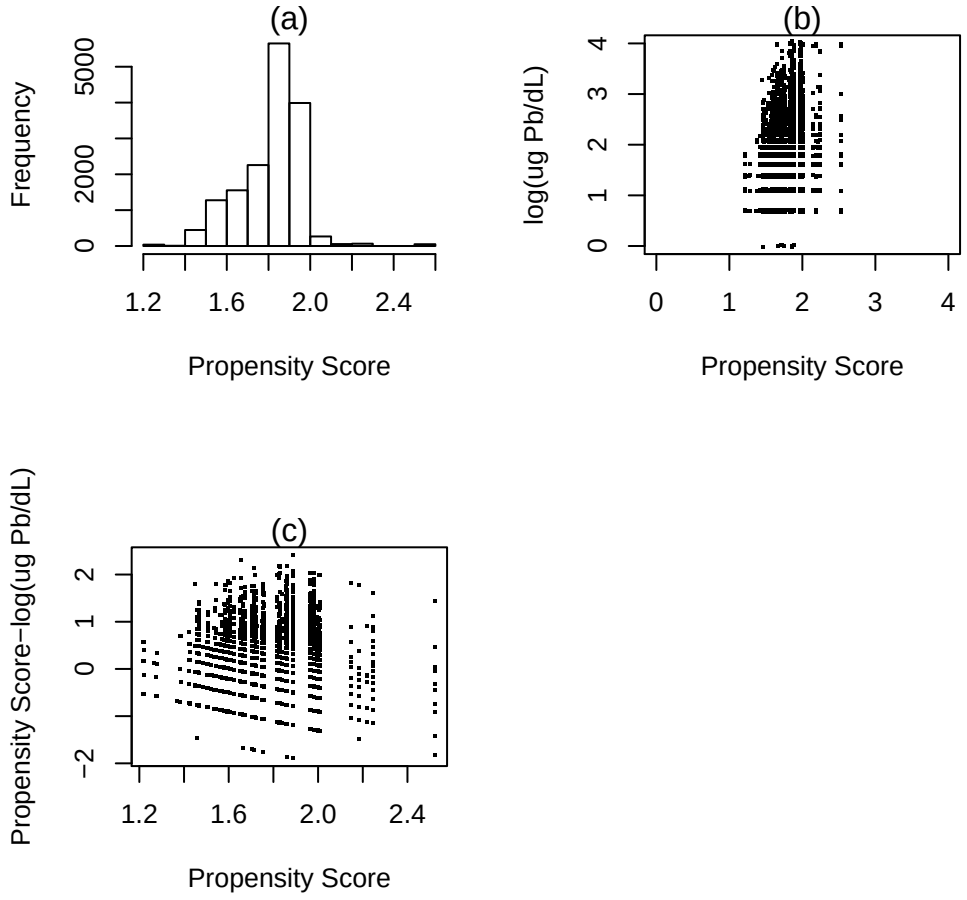

Figure 2: (a) Histogram of propensity scores; (b) Propensity scores vs. log blood lead level; (c) Residuals of propensity score model vs. log blood level level.

graders retain a significantly higher level of BLL (though reduced by a factor of three from the marginal difference between third and fifth-graders).

		Propensity Score Weight Classes							
SES measure	All	1	2	3	4	5	6	7	8
3rd grade (vs. 5th)	53.0	92.4	40.4	57.6	83.4	80.0	0	0	0
Male	55.5	45.7	49.0	51.8	60.0	58.8	55.1	5.3	100
African-American	89.8	89.5	67.5	68.3	79.0	98.9	100	94.7	61.3
ESL	17.1	22.1	23.4	31.5	32.7	7.8	6.6	94.7	61.3
Free-lunch-Eligible	80.1	0	23.3	29.5	97.9	98.6	100	91.1	38.7
Maternal Educ. $\leq$ HS	77.4	14.3	25.4	99.3	32.2	80.9	98.8	94.7	75.2
n	15652	552	863	2381	1843	5585	4066	225	137

Table 3: Covariate proportions, overall and by propensity score weight class.

		Propensity Score Weight Classes							
SES measure	All	1	2	3	4	5	6	7	8
3rd grade (vs. 5th)	<.001	.78	.91	.022	.82	.18	NA	NA	NA
Male	.069	.76	.22	.50	.17	.34	.24	.46	NA
African-American	<.001	.69	.95	.86	.74	.74	NA	.46	.18
ESL	<.001	.85	.68	.86	.84	.70	.41	.46	.18
Free-lunch-Eligible	<.001	NA	.70	.80	.76	.78	NA	.87	.18
Maternal Educ. $\leq$ HS	<.001	.26	.93	.53	.55	.12	.25	.46	.10
n	15652	552	863	2381	1843	5585	4066	225	137
W_j	—	.035	.055	.152	.118	.357	.260	.014	.009

Table 4: p-values associated with t-tests comparing log blood level levels with each covariate, overall and by propensity score weight class.

To proceed with the analysis, we first consider the degree of support for the BLL distribution within each of the propensity classes. Figure 3 shows that, across the propensity score classes, only log(BLL+1)s of between 1 and 3 (BLL of 2-19 μg/dL) have substantial support in all classes – i.e., meet the positivity criterion (Westreich and Cole 2010). So our analysis proceeds restricting the estimated effect of blood level level on reading and mathematics MEAP scores to between 2 and 19 μg/dL. This eliminates 1,227 students from the 15,652 in our analysis.

Next, we refit (4) within each propensity score class j, obtaining an estimate $\hat{\beta}_{pj}$ for $p = 0, ..., 4$, $j = 1, ..., 8$. We then compute $\hat{\beta}_p = \sum_{j=1}^{8} \hat{W}_j \hat{\beta}_{pj}$, where $\hat{W}_j$ is the sampled-estimated proportion of the population in each of the classes (see Table 3). The results are given in Table 4. In general the estimated causal effect of BLL appears to be somewhat reduced over that obtained by adjustment by covariates, although the effect remains significant.

Another advantage of the propensity score approach is the ability to estimate the overall probability of insufficient reading or mathematics proficiency after adjusting for covariates without having to fix a particular covariate pattern. Following (1), we obtain a point estimate of $P(Y(T) = 1)$ as $expit(\hat{\beta}_0 + \hat{\beta}_1 T + \sum_{q=1}^{3} \hat{\beta}_{q+1} S(T)_q)$, where $expit(x) =$

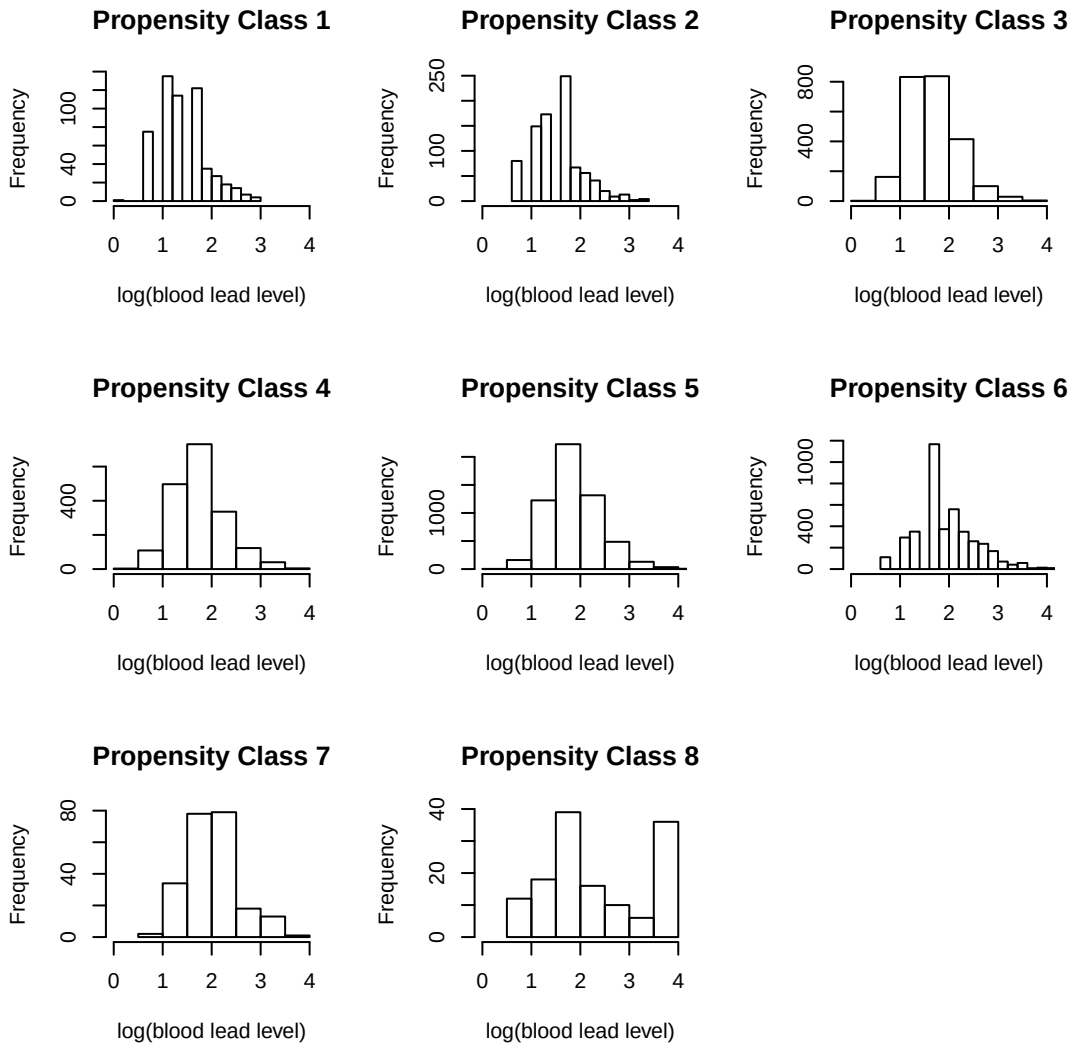

Figure 3: Histogram of log(blood lead level) (μg/dL) by propensity score class.

$\exp(x)/(1+\exp(x))$; variance estimates and associated 95% confidence intervals are obtained via the Delta Method. Figure 4 presents the predicted probability of insufficient reading proficiency at a given level of BLL between 2 and 19 μg/dL, unadjusted (black), adjusted to mean covariate values and fit using only children with observed BLL between 2 and 19 μg/dL (blue), and balanced with respect to observed covariates using propensity score modeling over the BLL range of 2 to 19 μg/dL (red). Unadjusted and propensity-score adjusted results are broadly similar, suggesting an approximately linear rise in probability through approximately 8-10 μg/dL, with approximately level risk up to 19 μg/dL. Failure to adjust for the confounding effects of gender, race, and SES factors leads to overestimation of the effect of lead on reading proficiency, although this overestimation is modest. Figure 5 presents the equivalent predicted probability of insufficient mathematics proficiency, again unadjusted, adjusted to mean values using only children with BLL between 2 and 19 μg/dL, and propensity score adjusted. Although the overall patterns are broadly similar, failing to restrict the BLL to a common support and failure to adjust for confounding leads to larger differences in the estimated effects of lead exposure on mathematics proficiency than on reading proficiency. Estimates using standard regression adjustments versus estimates using balancing via propensity scores differ little for predicting reading proficiency; for math, estimates of failure to achieve proficiency are somewhat higher for the propensity score method than the regression method in the 2-8 μg/dL range, with no clear difference above this level. For both reading and math, propensity score analysis suggests an increase in the probability of failing to reach sufficiency increasing from approximately one-fourth to one-third as blood lead increases from 3 μg/dL to 10 or more μg/dL.

For exposures above 19 μg/DL, we can restrict our analysis to those in weight propensity class 3 or greater (98.6% of students with exposures above 19 μg/DL were in weight propensity classes 3 or greater). Dropping the 2,308 students who were either 1) in the first or second propensity weight class, or 2) had an observed BLL of less than 2 μg/dL and fitting a propensity model (red) and comparing with an unrestricted, unadjusted model (black) for probability of insufficient reading proficiency up to maximum observed level of 70 μg/dL in Figure 6 shows some tendency of the risk of insufficient reading proficiency to increase in the propensity estimate, in contrast to an apparent threshold effect in the unadjusted model. For insufficient mathematics proficiency, both unadjusted and propensity estimates of risk increase with BLL, although the increase in risk is somewhat attenuated in the propensity model (Figure 7). The limited number of students exposed to such high levels of lead make precise inference difficult.

3.2 Non-bipartite Matching

Next we consider conditional logistic regression approach based on a non-bipartite matching of the propensity scores. Using the method of Lu et al. (2011), we develop an optimal matching, treating the absolute difference in the propensity score of BLL as the distance between two students. We are able to match all but 34 of the 15,652 students to within propensity scores (predicted log blood lead values) of .001 or less, yielding 7,809 pairs for analysis. Five of the six covariates – grade level, sex, race, free-lunch eligibility, and maternal education – are perfectly balanced across the two groups; only English as a second language was less than perfectly balanced, although it was very close: (17.0% in the first

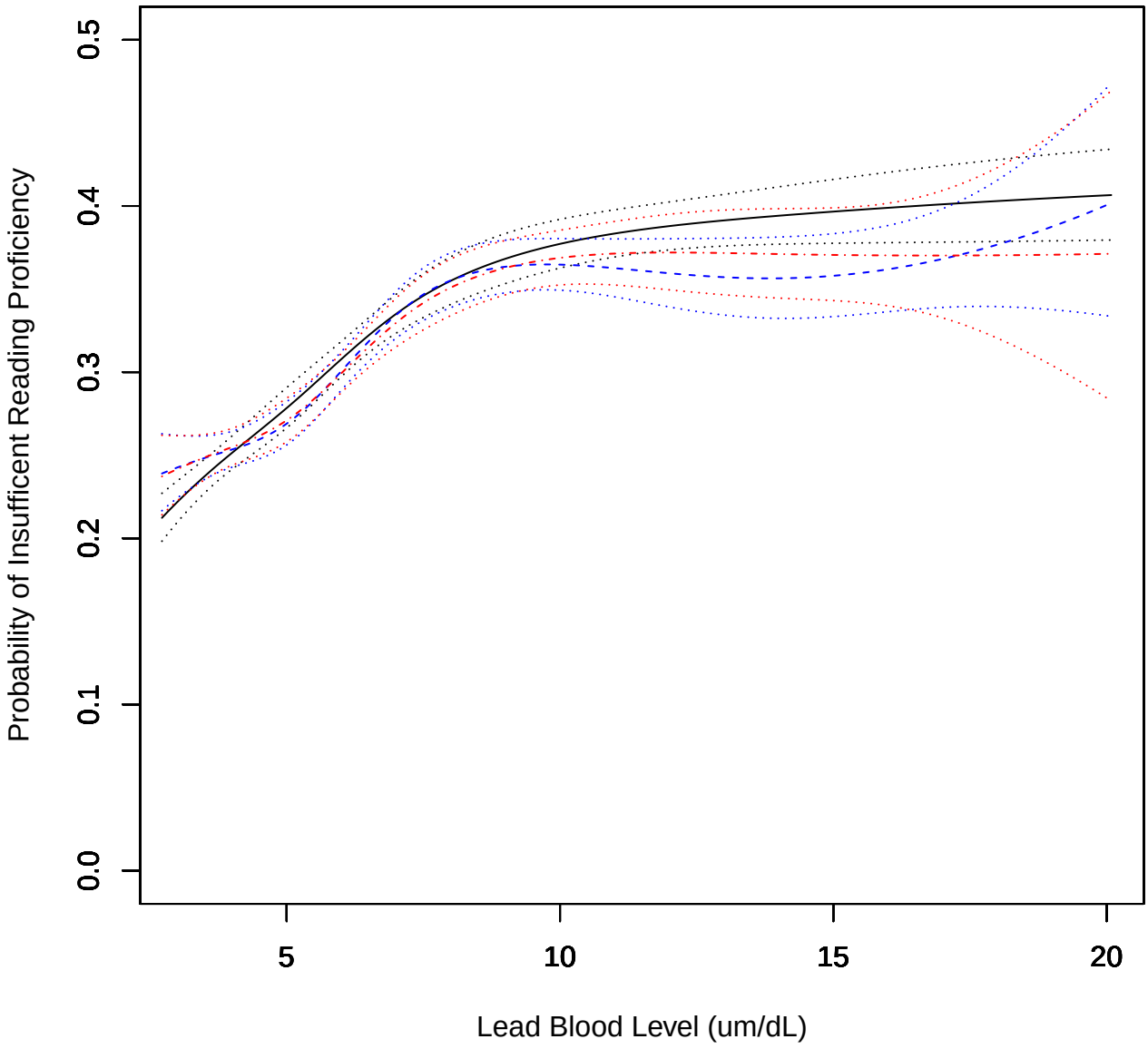

Figure 4: Predicted probability of insufficient reading proficiency for blood lead levels between 2 and 19 μg/dL: unadjusted using all students (solid black); regression-adjusted using only students with observed blood lead levels between 2 and 19 μg/dL, with confounders set to mean values (dashed blue); and estimated using propensity score modeling for levels between 2 and 19 μg/dL (dot-dash red). Dotted lines give 95% CIs.

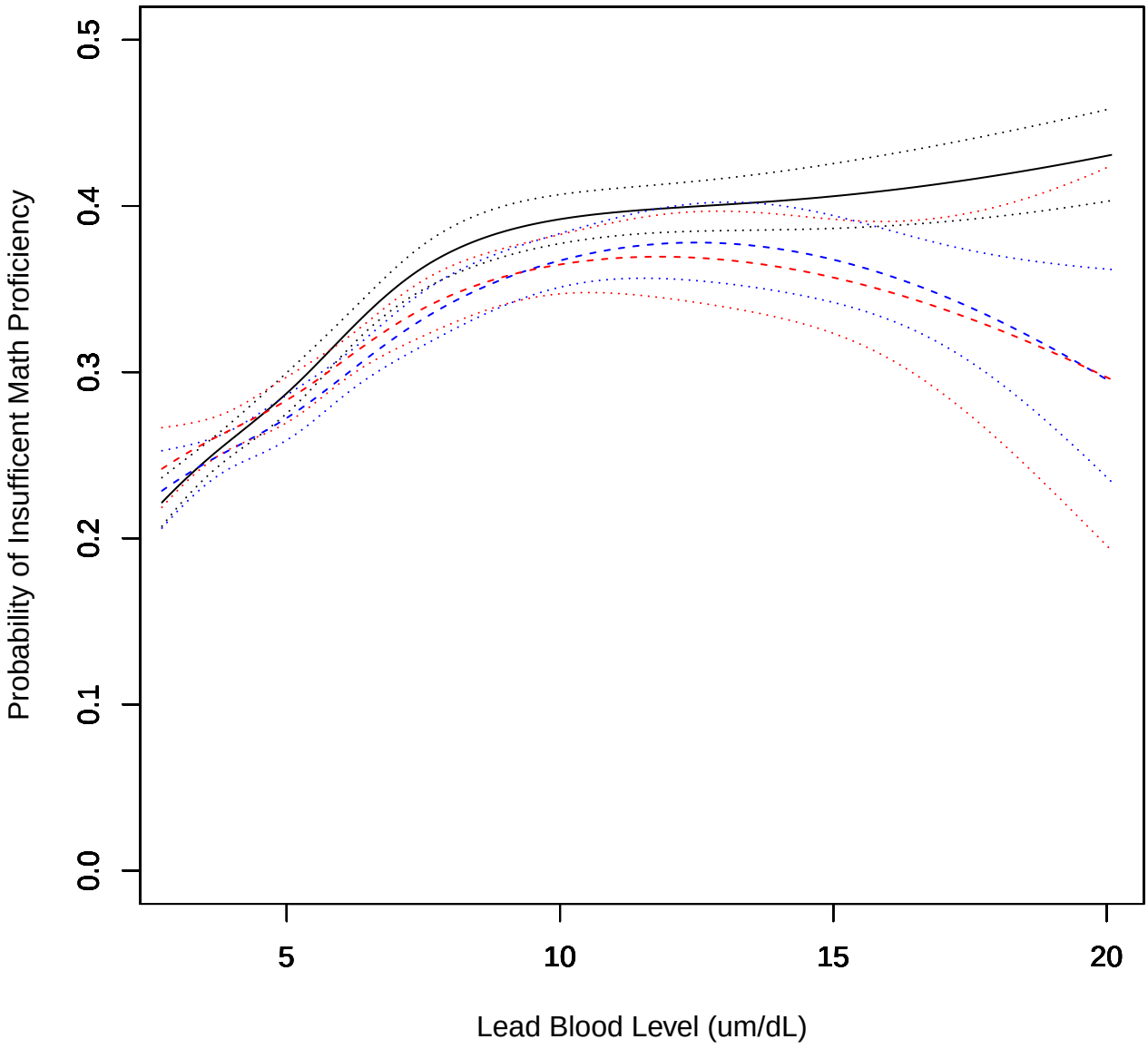

Figure 5: Predicted probability of insufficient mathematics proficiency for blood lead levels between 2 and 19 μg/dL: unadjusted using all students (solid black); regression-adjusted using only students with observed blood lead levels between 2 and 19 μg/dL (dashed blue), with confounders set to mean values; and estimated using propensity score modeling for levels between 2 and 19 μg/dL (dot dashed-red). Dotted lines give 95% CIs.

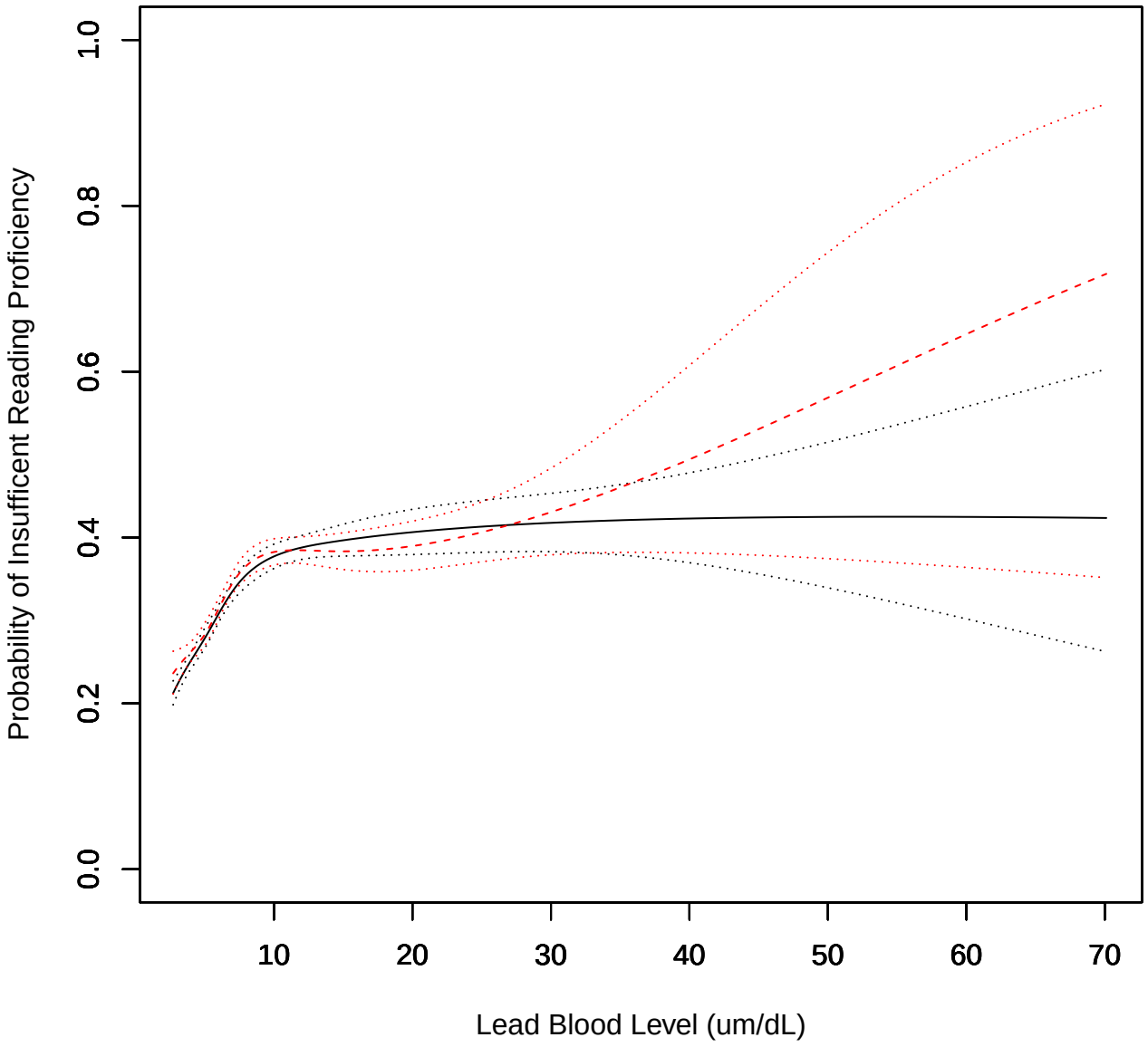

Figure 6: Predicted probability of insufficient reading proficiency for blood lead levels above 1 μg/dL: unadjusted using all students (solid black); and estimated using propensity score modeling for levels above 1 μg/dL estimated (dashed red) (among subjects with a non-trivial probability of having blood lead levels above 19 μg/dL based on regression modeling). Dotted lines give 95% CIs.

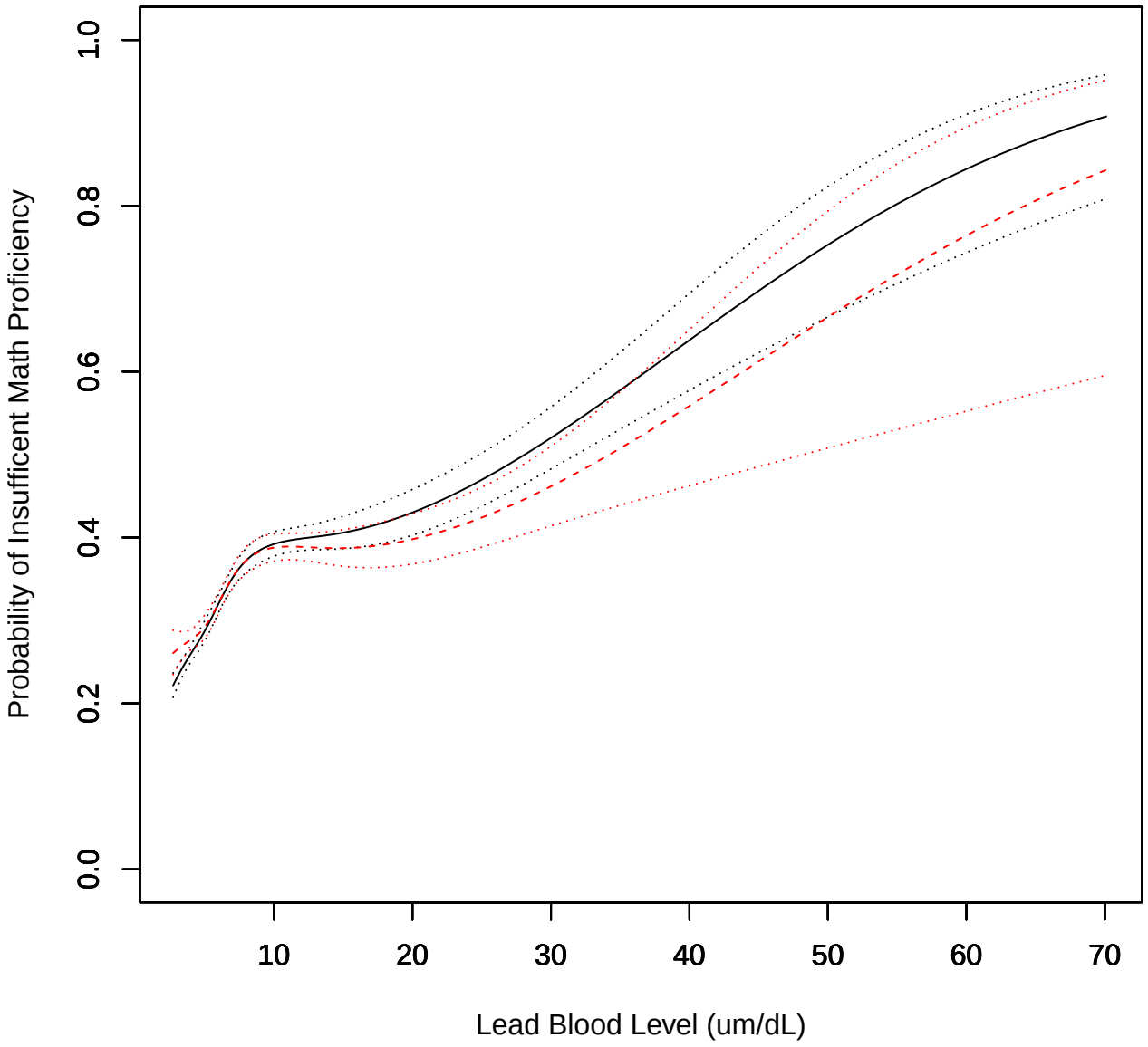

Figure 7: Predicted probability of insufficient mathematics proficiency for blood lead levels above 1 μg/dL: unadjusted using all students (solid black); and estimated using propensity score modeling for levels above 1 μg/dL (dashed red) (among subjects with a non-trivial probability of having blood lead levels above 19 μg/dL based on regression modeling). Dotted lines give 95% CIs.

group vs. 17.2% in the second group). As with the propensity score stratification, we need to consider the maximum difference in BLL among the pairs as the limit of support for considering causal inference effects. Figure 8 plots the log(BLL+1) levels for the pairs: as with the stratification approach, we see that almost no subjects with values of log(BLL+1) below 1 are paired with log(BLL+1) above 3, so, we retain this restriction in our analysis, limiting the matched pairs to the 6,636 pairs where both subjects had log(BLL+1) greater than 1 and less than 3. Table 4 provides the results of this conditional logistic regression analysis. They are similar to those obtained from the logistic regression averaged across the propensity strata for reading proficiency; the results differ somewhat for mathematics proficiency, with non-linear effects attenuated.

To increase power, we divided the students into 62 groups based on the rounding of the propensity score to the nearest one-thousandth. Dropping 4 of these groups due to their having 3 or fewer observations, we matched subjects based on $\mid \log(BLL+1)_i - \log(BLL+1)_j \mid^{-1}$ to maximize the difference in the observed BLL levels withing pairings based on BLL propensity. Figure 9 plots the log(BLL+1) levels for the BLL-discordant pairs. In contrast to the pair matching based only on propensity scores in Figure 8, we see nearly all pairings have a difference of at least .5 log(BLL+1); the mean absolute difference in log(BLL+1) has increased from .64 in the non-BLL-discordant pairs to .89 in the BLL-discordant pairs. Table 4 provides the results of a conditional logistic regression analysis based on this BLL-discordant maximization, again restricted to the 6,590 pairs where both subjects had log(BLL+1) greater than 1 and less than 3 for comparison. Results are very broadly similar to the analysis without maximizing separation; however, point estimates are generally larger in magnitude while standard errors are increased as well.

3.3 Bayesian Additive Regression Trees

Applying BART to the Detroit blood level data, we used the default priors associated with the "bart" function available in the R package BayesTree: a prior probability of 0.05, 0.55, 0.28, 0.09 and 0.03 for trees with 1, 2, 3, 4 and $\geq$5 terminal nodes respectively, $\mu_{ij} \sim N(0, (4\sqrt{m})^{-1})$, and $\sigma^2 \sim 3\lambda/\chi^2_3$, where m is the number of trees ($m = 200$ in default) and λ is chosen so that $P(\sigma < \hat{\sigma}) = .9$, where $\hat{\sigma}$ is based on a linear regression of the outcome on the covariates (Chipman et al. 2010). The set of counterfactual values $\mathcal{T}$ considered consists of log(BLL+1) starting at .5 and increasing by .1 through 4.5, while the comparison levels T^A and T^B are given by groupings of .5 of log(BLL+1) from .5 through 4.5. We obtained 5,000 draws from the MCMC chains after a burn-in of 100; running a second chain and using Gelman's $\sqrt{\hat{R}}$ measure (Gelman et al. 2003) gives a maximum value of 1.10 for the posterior draws of the counterfactual values of $Y(T), T \in \mathcal{T}$, sufficient for convergence. The results are given in the top row of Table 5 (for reading) and Table 6 (for math). For reading, the results are roughly consistent with those found through the other methods; approximately linear increases in risk from the lowest levels up through 10-15$\mu g/dL$, then a leveling in risk, with increases on the order of 20 percentage points in risk of failure to achieve MEAP reading proficiency. Results are similar for math, with somewhat stronger evidence for increases in risk of 20-40 percentage points around 30$\mu g/dL$, again consistent with the spline modeling results. For both reading and math, 7.39$\mu g/dL$ (log=2) appears to be the point at which the lead level yields statistically significant increases in risk of

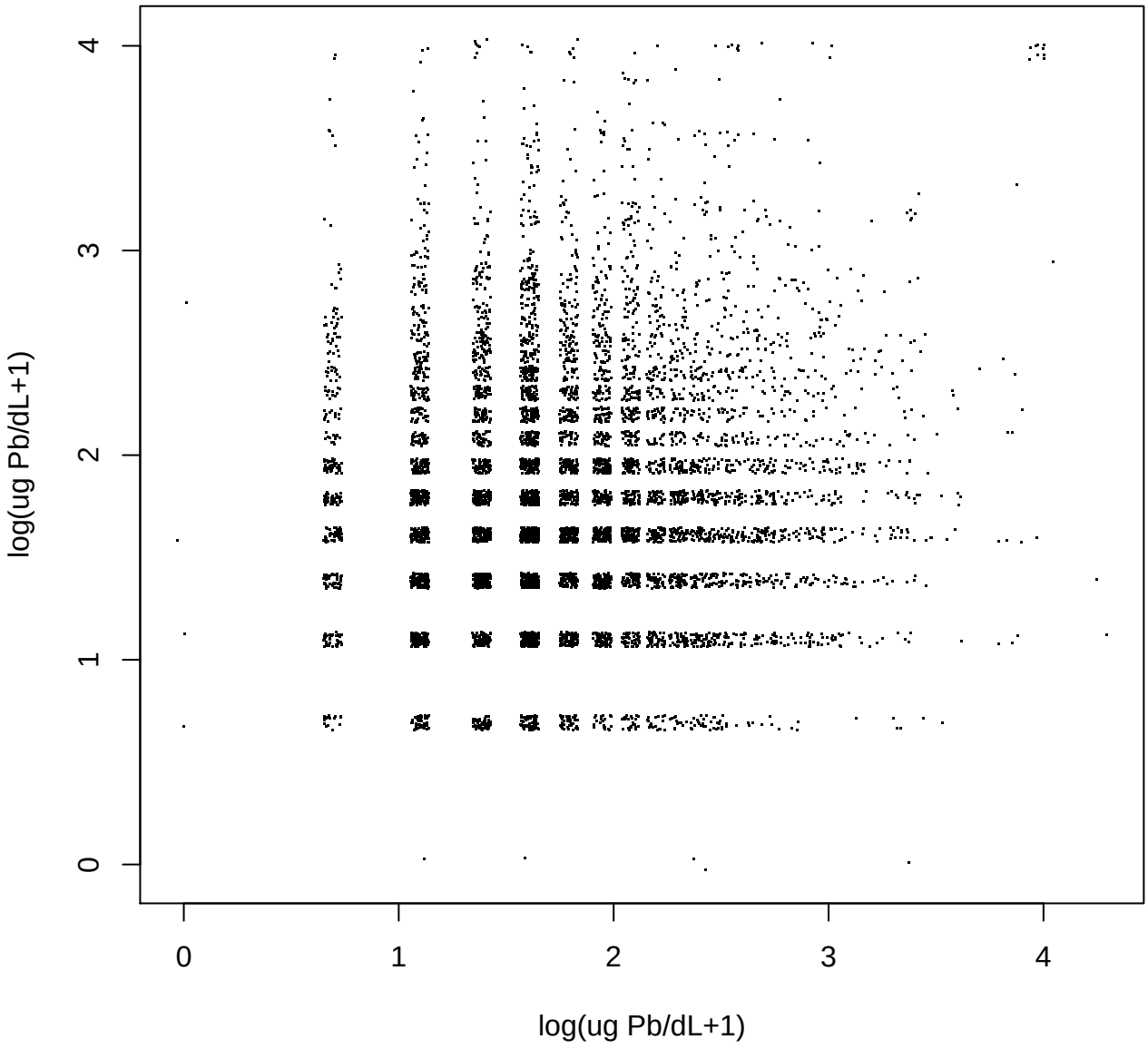

Figure 8: Log(μg/dL+1) for the matched pairings based on propensity scores (no blood lead level discrepancy maximization).

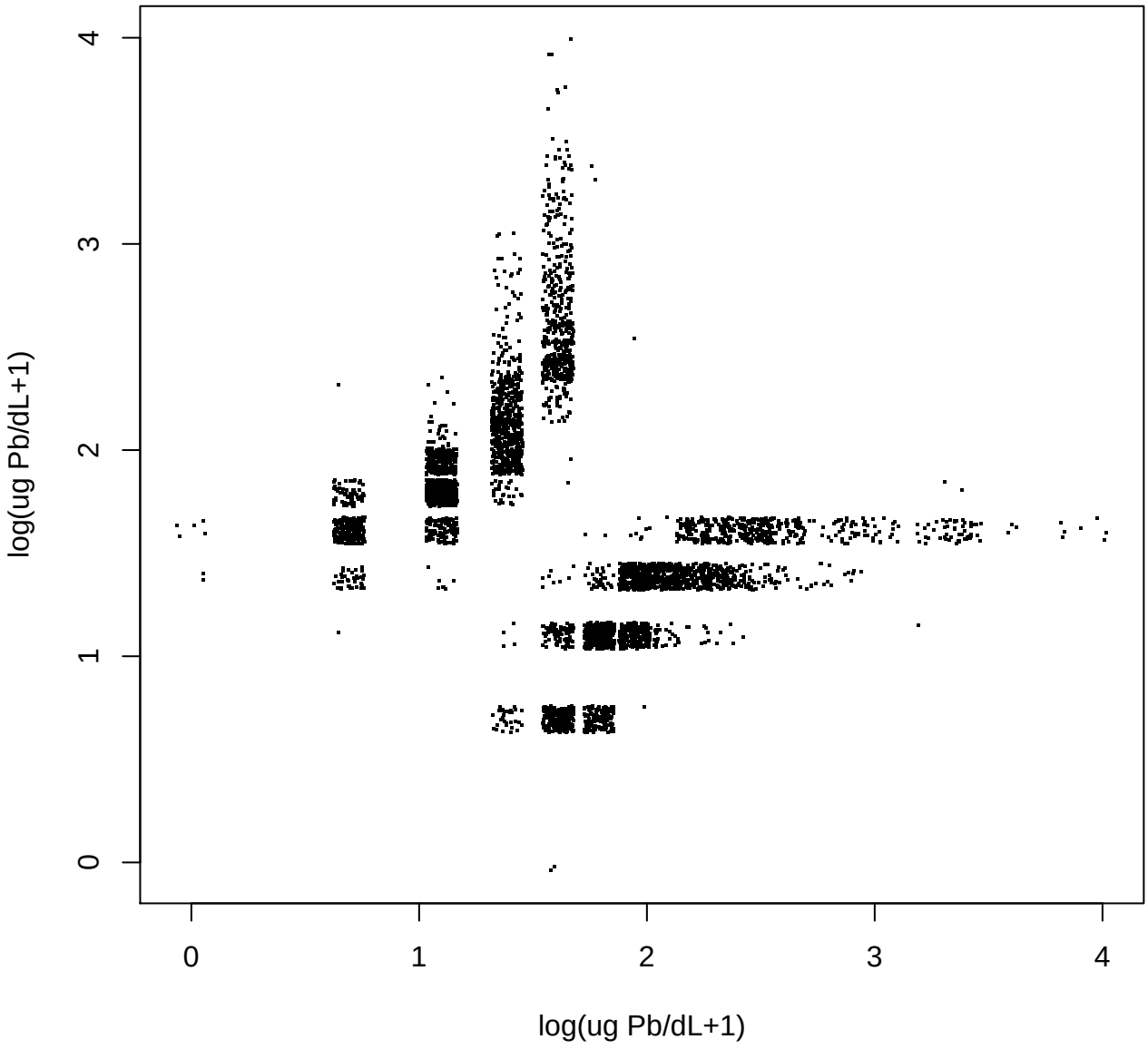

Figure 9: Log(μg/dL+1) for the matched pairings based on propensity scores (blood lead level-discrepancy maximized).

	T^B					
T^A	2.72	4.48	7.39	12.18	20.08	33.11
1.64	$.047_{-.073,.161}$	$.063_{-.060,.195}$	$.155_{.048,.262}$	$.163_{.056,.270}$	$.133_{.022,.250}$	$.321_{.170,.478}$
	$.012_{-.018,.039}$	$.017_{-.016,.050}$	$.041_{.012,.066}$	$.042_{.013,.072}$	$.027_{.000,.052}$	$.026_{.014,.039}$
	.292	.316	.316	.316	.269	.075
2.72		$.016_{-.083,.121}$	$.108_{.040,.179}$	$.116_{.037,.200}$	$.086_{-.001,.176}$	$.274_{.130,.421}$
		$.009_{-.034,.053}$	$.051_{.016,.086}$	$.055_{.017,.093}$	$.027_{.000,.057}$	$.028_{.013,.041}$
		.482	.542	.538	.338	.094
4.48			$.092_{.017,.175}$	$.100_{.019,.174}$	$.069_{-.014,.159}$	$.258_{.118,.406}$
			$.049_{.005,.097}$	$.053_{.007,.096}$	$.023_{-.007,.056}$	$.026_{.013,.039}$
			.612	.608	.381	.094
7.39				$.008_{-.047,.060}$	$-.022_{-.089,.043}$	$.166_{.039,.292}$
				$.008_{-.046,.059}$	$-.014_{-.042,.013}$	$.015_{.002,.027}$
				.984	.392	.094
12.18					$-.030_{-.100,.031}$	$.158_{.033,.289}$
					$-.019_{-.046,.008}$	$.013_{.001,.024}$
					.391	.094
20.09						$.188_{.059,.323}$
						$.017_{.005,.030}$
						.094

Table 5: Reading: Posterior mean of estimated change in probability $D(T^A, T^B)$ of failing to achieve MEAP proficiency resulting in a change of BLL from range T^A to T^B (95% posterior predictive intervals in subscript). Top row includes all observations; middle row includes observations with sufficient support; bottom row provides fraction of observations retained with sufficient support at the counterfactual dose level

failing to achieve proficiency. (Note that, given positivity, we have $d(a,b)+d(b,c) = d(a,c)$, but that this relationship does not hold at the highest dose levels given that only a small fraction of the sample is estimated to be able to have received both low and high levels of lead exposure given the observed covariates, and thus the positivity assumption begins to fail.)

Hill and Su (2013) also proposes to use BART to assess whether overlap is reasonable, based on the uncertainty in the posterior predictive distributions associated with the outcome at the observed versus the counterfactual dose level. Extending one of the options proposed by Hill and Su to the continuous setting, we exclude observations from the computation of $D^{(l)}(T^A, T^B)$ when the variance of the posterior predictive distribution of $E(Y_i^{(l)}(T))$ exceeds the variance of $E(Y_i^{(l)}(t_i))$ by a factor of $\chi^2_{1,1-q}$. The second row of Table 5 gives the results for $q = .05$ (excluding the counterfactual effects where there is at least a 95% chance that the posterior predictive variance of the outcome at the counterfactual dose level exceeds the posterior predictive variance of the outcome at the observed dose level, assuming the ratio variances follow a χ^2_1 distribution). The third row of Table 5 gives the fraction of the sample that is retained for the analysis. In general the effects of lead level are attenuated, particularly for the larger values of lead levels, although the associated credible intervals are also shrunk, so that results that were significant without

	T^B					
T^A	2.72	4.48	7.39	12.18	20.08	33.11
1.64	$.063_{-.067,.170}$	$.078_{-.046,.186}$	$.173_{.059,.276}$	$.168_{.057,.265}$	$.199_{.078,.304}$	$.214_{.075,.343}$
	$.033_{-.021,.080}$	$.045_{-.008,.090}$	$.083_{.035,.124}$	$.086_{.038,.125}$	$.066_{.025,.103}$	$.026_{.007,.043}$
	.490	.487	.485	.472	.353	.136
2.72		$.015_{-.074,.105}$	$.110_{.034,.197}$	$.105_{.037,.181}$	$.136_{.054,.238}$	$.151_{.039,.263}$
		$.013_{-.053,.077}$	$.081_{.025,.146}$	$.077_{.029,.133}$	$.067_{.021,.125}$	$.020_{.003,.037}$
		.730	.729	.715	.552	.155
4.48			$.094_{.022,.181}$	$.090_{.019,.165}$	$.120_{.041,.210}$	$.136_{.026,.256}$
			$.079_{.020,.148}$	$.072_{.016,.130}$	$.063_{.019,.113}$	$.018_{.001,.036}$
			.816	.800	.561	.155
7.39				$-.005_{-.069,.056}$	$.026_{-.048,.105}$	$.041_{-.068,.155}$
				$-.004_{-.059,.048}$	$.010_{-.034,.059}$	$.004_{-.013,.020}$
				.853	.586	.155
12.18					$.031_{-.037,.103}$	$.046_{-.054,.146}$
					$.016_{-.023,.059}$	$.004_{-.011,.019}$
					.585	.155
20.09						$.016_{-.099,.117}$
						$.001_{-.016,.017}$
						.155

Table 6: Math: Posterior mean of estimated change in probability $D(T^A, T^B)$ of failing to achieve MEAP proficiency resulting in a change of BLL from range T^A to T^B (95% posterior predictive intervals in subscript). Top row includes all observations; middle row includes observations with sufficient support; bottom row provides fraction of observations retained with sufficient support at the counterfactual dose level

this elimination remain significant. In particular, the large effects estimated on reading proficiency between very low and very high BLLs are substantially reduced when subjects unlikely to be observed at both dose levels are removed, suggesting that these very large impacts are due to extrapolation to subjects who in actuality would have been unlikely to have been able to receive both levels. (Note that this does not preclude the possibility of such effects, but does require the proposed model to be accurate in its extrapolation.)

Of the $n \times 41$=641,732 combinations of observed and counterfactual BLL levels considered, 39% reach the exclusion cutoff level of $\chi^2_{1,.95} = 3.841$ for the reading scores. Exclusion rates are slightly higher (47%) for the mathematics scores, although the impact of the exclusion is similar. Tables 5 and 6 show that a majority of subjects have stable counterfactual estimates of blood lead levels only in the range of 2-20 $\mu g/dL$, consistent with the restrictions we provided in Figures 4 and 5.

4. Discussion

Among third and fifth graders in Detroit Public Schools, those who were in 5th grade, were African-American, spoke English as a Second Language (ESL) or were free-lunch eligible, or had mothers with a less-than high school education were all more likely to have higher blood-level levels as toddlers or young children, and were less likely to test as proficient in reading or mathematics in the MEAP test exams, the standardized tests given to all 3rd and 5th graders (among others) in the State of Michigan. When Zhang et al. (2013) considered the association between early childhood lead exposure and reading/math proficiency as measured by the MEAP exams, they adjusted for these factors in their logistic regression analyses.

However, correct adjustment for these factors assumed correct formulation in the model (they were included as independent logit-linear predictors without interactions) and that children at each level of the potential confounders had the entire range of BLL represented. We determined this was not the case for child age (virtually all children with low BLL were third graders, while those with the highest levels of BLL were without exception 5th graders), and very few children with low BLL were ESL, free-lunch eligible, or had mothers with low levels of education. Propensity score analysis removes the need to explicitly model the confounder as a function of the outcome mean, and allows truncation of subjects who would likely have been eligible to receive only one extreme or another of the dosage level (here BLL). Methods for propensity scores with continuous outcomes have only recently been developed, and their use is limited. We consider three approaches here: a "standard" approach that averages over propensity strata, excluding levels of BLL that are not supported in all strata; a non-bipartite matching strategy that matches subjects based on estimated BLL propensity score, and a Bayesian additive regression tree (BART) approach that provides a robust prediction of the counterfactual values of proficiency and thus computes the causal effects of BLL exposure directly, again restricting to reasonable areas of BLL exposure differences based on minimal levels of posterior predictive variability.

We found that, while the BLL range in the sample was from 1 to over 80 $\mu g/dL$ in their early childhood, only children with ranges from less than 2 to 19 $\mu g/dL$ were observed across the entire range of covariates. Our analysis indicated that, while children still had rapid increase in risk of insufficient reading and/or mathematics proficiency from 2 to 10

$\mu g/dL$, this impact was somewhat attenuated when restricted to children who could have been exposed to the full range of BLL exposures from 2 to 19 $\mu g/dL$. A similar analysis using non-bipartite matched propensity scores yielded similar results. The BART analysis also concluded broadly similar effects on the change of BLL on the risk of non-proficiency in reading and/or math; however, it indicated 1) that increases past 10-15 $\mu g/dL$ in BLL may have had little effect on reading and mathematics proficiency, and 2) that the lack of covariate overlap was important for projected effects of large differences in BLL on reading and mathematics proficiency. Nonetheless, this work provides evidence that recent CDC calls (CDC 2012) to reduce BLL to 5 μg/DL or less are warranted, as substantial and statistically significant increases in the risk of not-proficient reading and mathematics tests scores were observed above this level in all analyses considered.

Of course, the three methods we consider for assessing the causal effect of lead exposure on reading or mathematics proficiency are not the only ones available. Other options include variations on the direct model of (2), such as a standard adjusted logistic regression model (replacing $g(e_\theta(x_i), \gamma)$ with $\gamma' x_i$) or estimation of the counterfactual based on a low-degree polynomial of $e_\theta(x_i)$ (e.g., $g(e_\theta(x_i), \gamma) = \gamma_0 + \gamma_1 e_\theta(x_i) + \gamma_2 e_\theta(x_i)^2$) rather that the BART model.

All of the results in this manuscript rely on the strong ignorability assumption, that is, there are no unmeasured confounders. However, our set of available confounders, while important (e.g., mother's educational level), was quite small – only six. Hence our conclusions must be tempered by the likely possibility that unobserved confounders may be present that explain yet more of the lead exposure-school testing relationship.

We believe that the practical development of propensity score methods for continuous or, more generally, non-binary exposure remains a rich field for exploration. One issue that we did not consider here is the development of the propensity score itself. We used a simple linear model, which implies, if not a normally-distributed exposure, at least a reasonably symmetric and not too-heavy-tailed continuous distribution, as was the case here with the log of blood level level transformation. Alternatives such as kernel density regression estimators (Hardel 1990), or exponential titled regression models (Rathouz and Gao 2009) might be considered as alternative propensity score estimators in other settings.

Acknowledgements

The authors would like to thank Harolyn Baker, Margaret Tufts, and Randall Raymond for providing access to the Michigan Department of Community Health Childhood Lead Poisoning Prevention and Control Program blood lead surveillance data and the Michigan Educational Assessment Program standardized test scores to be used for this work.

References

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

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Centers for Disease Control and Prevention (2012). CDC response to advisory committee on childhood lead poisoning prevention recommendations in Low Level Lead Exposure Harms Children: A Renewed Call of Primary Prevention. Available at:

http://www.cdc.gov/nceh/lead/ACCLPP/ CDC_Response_Lead_Exposure_Recs.pdf. Accessed June 1, 2012.

Chipman, H., George, E. and McCulloch, R. (1998). Bayesian CART model search. *Journal of the American Statistical Association* 93:935-948.

Chipman, H., George, E. and McCulloch, R. (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics* 4:266-298.

Derigs, U. (1988). Solving nonbipartite matching problems via shortest path techniques. *Annals of Operations Research* 13:225-261.

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

Gelman, A., Carlin, J.B., Stern, H.S.and Rubin, D.B. (2003). *Bayesian Data Analysis second edition*. London: CRC press, 2003.

Hardel, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press, Cambridge.

Hill, J.L. (2011). Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20:217-240.

Hill, J.L and Su, Y.S.S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *Annals of Applied Statistics* 7:1386-1420.

Imai, K. and Van Dyk, D.A. (2004). Causal Inference with General Treatment Regimes: Generalizing the Propensity Score. *Journal of the American Statistical Association* 99:854-866.

Joffe, M.M. and Rosenbaum, P.R. (1999). Propensity scores. *American Journal of Epidemiology* 150:327-333.

Lu, B., Greevy, R., Xu, X. and Beck, C. (2011). Optimal Nonbipartite Matching and Its Statistical Applications. *The American Statistician* 65:21-30.

Lu, B., Zanutto, E., Hornik, R. and Rosenbaum, P.R. (2001). Matching with doses in an observational study of a median campaign against drug abuse. *Journal of the American Statistical Association* 96:1245-53.

Rathouz, P.J. and Gao, L.P. (2009). Generalized linear models with unspecified reference distribution. *Biostatistics* 10:205-218.

Rubin, D.B. (1973). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, 29:185203.

Rubin, D.B. (1980). Randomization analysis of experimental data: the Fisher randomization test. *Journal of the American Statistical Association* 75:591-593.

Westreich, D. and Cole, S.R. (2010). Invited commentary: Positivity in practice. *American Journal of Epidemiology*, 171:674-677.

Zhang, N., Baker, H.W., Tufts, M., Raymond, R.E., Salihu, H. and Elliott, M.R. (2013). Early Childhood Lead Exposure and Academic Achievement: Evidence from Detroit Public Schools (2008-2010). *American Journal of Public Health*, 103:e72-e77.