Supplementary Materials for

Children Drinking Private Well Water Have Higher Blood Lead Than Those with City Water

Jacqueline MacDonald Gibson, John M. MacDonald, Michael Fisher, Xiwei Chen, Aralia Pawlick, Philip J. Cook
Correspondence to: jacmgibs@iu.edu

This PDF file includes:

- Data matching methods
- Summary statistics
- Missing values imputation methods
- Statistical analysis methods
- Evaluation of instrument validity
- Regression model output
- Tables S1 to S8
- Information about access to original data

North Carolina Department of Public Safety Matching Algorithm

The North Carolina (NC) Department of Public Safety (DPS) provided data on whether there were any juvenile justice complaints filed anywhere in NC against any of the 59,483 children included in the data set used in this analysis. To compile these data, we provided DPS with a database including each child's first name, last name, date of birth, and address at the time of Pb testing. DPS then searched all of its records through the end of 2019 to see whether there were complaints filed against any of these children in the juvenile justice system. For children with records, DPS provided data on the number of complaints, types of complaints (e.g., serious, violent), and date of first offense.

DPS used a multi-step matching process to identify children with complaint records. In the first step, matching was done on the first initial of the first name, the first three letters of the last name, and date of birth, resulting in 707 potential matches. This initial match step sought to ensure that children with juvenile records were not missed due to spelling errors in the first or last name in either data set.

The next step was to confirm matches. First, children for whom first names and last names matched exactly in the two data sets (n=435) were identified. The remaining 272 juvenile records were then searched manually for differences in spelling of first or last names. First names were checked first. When the first name in one data set differed from that in the other only in spelling or in one or two letters (for example, George in one source misspelled as Georg in the other), birth dates were checked, and the records were considered to match if both data sources listed the same birthdate. This yielded an additional 89 matches. The same approach was then used for last names, flagging 73 matches. After elimination of duplicates, a total of 568 unique matches were identified.

As mentioned in the main text, the analysis presented here includes only the children in our data set of 59,483 who reached age 14 by December 31, 2019 (n=17,858) and for whom we were able to obtain race and residential address history information from the NC Education Research Data Center (NCERDC). The final data set that included 13,647 individual children, of whom 423 had recorded complaints in the NC juvenile justice system.

NC Education Research Data Center Matching Algorithm

The NCERDC provided data on the race of and residential address histories for the children in our data set. NCERDC performed the linkage between our data set and NC statewide public and charter school records using the following algorithm:

1. Exact match on school district, that is local education agency (LEA), or county and the child's first name, last name, and date of birth
2. Exact match on first and last names and date of birth
3. Exact match on LEA or county and first and last names, but allow for mistakes in one of day, month, or year of birth
4. Exact match on LEA or county, last name, and date of birth, allowing for close first name or nickname
5. Exact match on LEA or county, first name, and date of birth, allowing for close last name
6. Exact match on last name and date of birth, allowing for close first name or nickname
7. Exact match on first name, date of birth, allow for close last name

8. Exact match on first and last name, but allow for mistakes in one of day, month, or year of birth
9. Exact match on first and last name

NCERDC was able to locate matching records for 13,647 of the 17,858 children who were 14 by the end of 2019. Table S1 shows the percentage of children matched at each step.

Table S1. NCERDC match results

Match Step	Percentage of Matches
1	68.63
2	9.49
3	3.93
4	13.63
5	1.51
6	0.57
7	1.51
8	0.73
9	0.00

Evaluation of Differences Between Children with and without School Records

Because NCERDC was not able to match all 17,858 children to records in the public school system, we checked for demographic differences among matched and unmatched children. All children who were matched are tracked with a unique "mastid" that remains unchanged throughout their years of schooling.

Table S2 summarizes the results of this comparison. As shown in the table, there were no significant differences among children with and without mastids on the key exposure variables of interest in this analysis, specifically

- blood Pb levels
- water source
- year of home construction (a proxy for exposure to lead paint and dust)

The two samples did differ significantly on demographic variables:

- Proportion black: Those without school records lived in census block groups with a lower percentage of black residents than those with school records (25% vs. 28%).
- Home value: Those without school records lived in homes that, on average, had slightly higher values ($205,000 vs. $185,000).
- Gender: Those without school records are more likely to be female than those with school records (51.7% vs. 49.8%).

For other demographic variables, although differences were significant, they were very small (e.g., the average birth year for those without records was 2002.4 vs. 2002.6 for those with school records). We address these differences by including controls for demographic variables in our regression models.

The proportion of children with juvenile complaint records differed for the two samples. Those without matching school records were less likely to have records than those with school records (1.3% vs. 3.1%).

Table S2. Comparison of children with and without public school records in NC

Variable	Group (for Categorical Variables)	With Mastid	Without Mastid	*p*
N		13,647	4,211	
Birth year (mean(SD))		2002.62 (1.99)	2002.36 (2.10)	<0.001
Proportion black in census block at blood test year (mean(SD))		0.278 (0.222)	0.253 (0.210)	<0.001
Blood test year (mean(SD))		2004.58 (1.88)	2004.42 (1.84)	<0.001
Log of blood lead (mean(SD))		0.678 (0.605)	0.692 (0.611)	0.314
Year of construction of child's residence at blood test year (mean(SD))		1983.74 (21.2)	1983.07 (21.8)	0.085
Child's age (months) at time of blood test (mean(SD))		23.58 (17.1)	24.94 (17.7)	<0.001
Log of value ($) of child's residence at blood test year (mean(SD))		12.13 (0.601)	12.23 (0.668)	<0.001
Water source (N (%))	Community water supply	7,628 (83.6)	2,178 (84.2)	0.503
	Private well	1,492 (16.4)	409 (15.8)	
Gender (N (%))	Female	6,794 (49.8)	2,176 (51.7)	0.005
	Male	6,736 (49.4)	1,983 (47.1)	
	NA	117 (0.857)	52 (1.24)	
Season of blood test (N (%))	Winter	3,352 (24.6)	1,006 (23.9)	0.556
	Spring	3,767 (27.6)	1,207 (28.7)	
	Summer	3,364 (24.7)	1,037 (24.6)	
	Fall	3,164 (23.2)	961 (22.8)	
Juvenile complaint record (N (%))	No complaint	13,224 (96.9)	4,156 (98.7)	<0.001
	At least one complaint	423 (3.10)	55 (1.31)	
Record of serious juvenile offense (N(%))	No serious complaint	13,452 (98.6)	4,185 (99.4)	<0.001
	At least one serious complaint	195 (1.43)	26 (0.617)	

NOTES: Mann-Whitney tests were performed for birth year, proportion black in census block, blood test year, log of blood Pb, year of construction, and age since these variables are not normally distributed. Independent *t*-test was performed for "log_home_value". Chi-square tests were performed for categorical variables.

The slightly higher value of the homes in which those without school records live and lower proportion of Black residents in their neighborhoods may suggest that some of the children without records in the NCERDC system may have attended private schools.

Table S3. Summary statistics, paired data set

Variable	All Children on Private Wells	Community Water	Private Wells Matched to Community	*p* (All Wells vs. Matched Wells)	*p* (Matched Wells vs. Community)
N	2,163	1,949	1,949		
Characteristics of Child					
Gender (*N* (%))					1.00
Female	1,078 (50.4)	975 (50.2)	975 (50.2)	0.878	
Male	1,060 (49.6)	968 (49.8)	968 (49.8)		
Race (*N* (%))					1.00
All other	1,714 (79.2)	1,579 (81.0)	1,579 (81.0)	0.155	
Non-Hispanic Black	449 (20.8)	370 (19.0)	370 (19.0)		
Age at time of Pb test, months (*N* (%))					1.00
0 to 8	14 (0.647)	2 (0.103)	2 (0.103)	0.002	
9 to 14	1,218 (56.3)	1,184 (60.7)	1,184 (60.7)		
15 to 19	147 (6.796)	101 (5.18)	101 (5.18)		
20 to 29	327 (15.1)	266 (13.6)	266 (13.6)		
>30	457 (21.1)	396 (20.3)	396 (20.3)		
Age indicator (mean (SD))	0.803 (0.304)	0.794 (0.311)	0.802 (0.303)	0.788	0.434
Year of blood test (mean (SD))	2005 (1.86)	2005 (1.77)	2005 (1.77)	0.873	1.00
Blood draw type (*N* (%))					0.982
Capillary	210 (10.7)	180 (10.1)	180 (10.1)	0.567	
Venous	1,750 (89.3)	1,599 (89.9)	1,595 (89.9)		
Child Outcomes					
Blood Pb, µg/dL (mean (SD))	2.52 (1.65)	2.31 (1.95)	2.48 (1.54)	0.567	<0.001
Blood Pb ≥ 5 µg/dL *(N* (%))	235 (10.9)	151 (7.75)	201 (10.3)	0.566	0.005
Any juvenile complaint (*N* (%))	64 (2.96)	47 (2.41)	53 (2.72)	0.645	0.543
# juvenile complaints (mean (SD))	0.144 (1.39)	0.135 (1.62)	0.143 (1.44)	0.645	0.536
Any serious complaint (*N* (%))	31 (1.43)	24 (1.23)	26 (1.33)	0.786	0.776
# serious complaints (mean (SD))	0.0587 (0.730)	0.0370 (0.772)	0.0595 (0.759)	0.787	0.769
Characteristics of Child's Home at Time of Blood Test					
Home value (mean (SD))	$226,575 ($257,905)	$214,972 ($180,590)	$224,741 ($251,838)	0.716	0.594
Built before 1978 (*N* (%))	598 (31.0)	544 (30.7)	544 (30.7)	0.852	1.00
Year built (*N* (%))					1.00
Before 1950	124 (6.43)	95 (5.36)	95 (5.36)	0.717	
1950-1977	474 (24.6)	449 (25.4)	449 (25.4)		
1978-1987	283 (14.7)	256 (14.5)	256 (14.5)		
1988-1997	482 (25.0)	448 (25.3)	448 (25.3)		
1998-2002	342 (17.7)	330 (18.6)	330 (18.6)		
2003 or later	224 (11.6)	193 (10.9)	193 (10.9)		
Characteristics of Child's Census Block Group, Blood Test Year					
Median income (mean (SD))	$82,795 ($32,866)	$71,028 ($30,981)	$83102 ($32,801)	0.738	<0.001
% Black (mean (SD))	21.0% (16.8%)	27.6% (23.6%)	20.8% (16.7%)	0.808	<0.001
Characteristics of Child's Census Block Group, Age 14					
Median income (mean (SD))	$77,134 ($32,833)	$75,877 ($32,753)	$77,635 ($33,056)	0.673	0.115
% Black (mean (SD))	22.5% (19.4%)	24.6% (21.6%)	22.2% (19.2%)	0.756	0.310

Missing Values Imputation

A *k* nearest neighbors (KNN) approach was applied to impute the missing values for key independent variables, including water source, blood test year, child's age at the time of blood testing, age indicator, gender, year of construction of child's residence at blood test year, proportion black in census block at blood test year, log of value ($) of child's residence at blood test year, blood draw type, birth year, proportion black in census block at blood test year and 14th birthday year, and median household income in census block group at blood test year and at 14th birthday year. Imputed values were computed with the *knnImputation* function in the "DMwR" package.

A smoothing strategy was applied to the response variable (blood Pb concentration), because blood Pb was reported only to the nearest integer, and leaving these data in integer format resulted in biased regression estimates. Therefore, integer estimates of blood Pb for measurements ≤ 5 µg/dL were converted to continuous values using the following approach:

Step 1. Mean and standard deviation (SD) for the log of blood Pb while accounting for left censoring of data were calculated using the *enormCensored* function in the "EnvStats" package.

Step 2. Integer-level observations were selected into six groups: 1, 2, 3, 4, 5, and >5 µg/dL. Each group's corresponding percentile range in the full distribution of blood Pb was recorded.

Step 3. Integer values of observations were smoothed using normally distributed random values with the mean and SD computed from Step 1, along with randomly selected probabilities uniformly distributed within the percentile range of each group. (Function *runif* and *qnorm* in the "stats" package were used in this step.)

This process was repeated 40 times to yield 40 data sets with imputed values, and all models were run on all 40 data sets. Reported regression coefficients represent averages across these 40 data sets.

Statistical Analysis Methods

First-stage models were fitted with the *glm* function in the "stats" package in R. Second-stage models were fitted with the *ivglm* function in the "ivtools" package.

Evaluation of Instrument Validity

Table S4. Test of exclusion restriction

Variable	Coefficient (n=13,580)[a]	Odds Ratio	Standard Error	*p*
Well water	0.179	1.20	0.142	0.21
Blood Pb (natural log)	0.195	1.22	0.0641	<0.01
Age indicator	1.60	4.93	0.243	
Male sex at birth (reference=female)	0.937	2.55	0.111	<0.001
Black race (reference=all other)	1.29	3.65	0.117	<0.001
Census block group median income (natural log) at age 14	-0.969	0.379	0.133	<0.001
Census block group % Black at age 14	0.319	1.38	0.256	0.21

[a]Coefficients from the following model (McFadden's R^2=0.145):

$$log\left(\frac{P(Delinquent)_i}{1-P(Delinquent)_i}\right) = \omega(WaterSource_i) + \beta(\log\,(BloodPb_i) + \vec{\gamma}\overrightarrow{C_\iota}+\varepsilon_i\text{Mc}$$

Table S5. Statistics for *J* tests of instrument exogeneity

Outcome	Model 1: All Children (n=13,580)[a]	Model 2: Children on Private Wells Matched to Children on Community Water (n=3,898)[a,b]
Any delinquency	*J*(df=5)=6.95, *p*=0.225	*J*(df=5)=5.35, *p*=0.375
Serious delinquency	*J*(df=5)=8.98, *p*=0.110	*J*(df=5)=3.49, *p*=0.625

Regression Output for Model 1

Table S6. Model 1, stage 1 (influence of water source and other variables on blood lead in early childhood)[a]

Variable	**Coefficient (n=13,580)[b]**	**Odds Ratio**	**Standard Error**	***p***
Private well water (reference=community water)	0.102	1.11	0.0208	<0.001
Year of blood test	-0.117	0.889	0.00554	<0.001
Home value (natural log)	-0.083	0.921	0.0173	<0.001
Construction year of residence during early childhood (reference=before 1950)				
1950-1977	-0.141	0.869	0.0334	<0.001
1978-1987	-0.201	0.818	0.0353	<0.001
1988-1997	-0.241	0.786	0.0347	<0.001
1998-2002	-0.329	0.720	0.0347	<0.001
2003 or later	-0.248	0.781	0.0379	<0.001
Age at time of blood test, months (reference=<9 months)				
9-14	0.105	1.11	0.0828	0.206
15-19	0.299	1.35	0.0865	<0.001
20-29	0.314	1.37	0.0845	<0.001
>30	0.279	1.32	0.0845	<0.001
Blood draw type venous (reference=capillary)	0.180	1.20	0.0234	<0.001
Black race (reference=all other races)	0.236	1.27	0.0177	<0.001
Male sex at birth (reference=female)	0.020	1.02	0.0137	0.140
Census block group median income (natural log) at blood test	-0.043	0.958	0.0267	0.106
Census block group % Black at time of blood test	0.026	1.03	0.0667	0.696
Census block group median income (natural log) at age 14	-0.027	0.973	0.0220	0.221
Census block group % Black at age 14	0.174	1.19	0.0476	<0.001
Fraction of age 16 reached at end of data collection	-0.119	0.888	0.0320	<0.001
Zip code (reference=27501)				
27502	-0.367	0.693	0.361	0.310
27511	-0.493	0.611	0.360	0.171
27513	-0.417	0.659	0.360	0.247
27518	-0.422	0.656	0.368	0.252
27519	-0.423	0.655	0.365	0.246
27520	-0.023	0.977	0.424	0.956
27522	-0.279	0.757	0.406	0.492
27523	-0.282	0.754	0.376	0.453
27526	-0.323	0.724	0.359	0.367
27529	-0.446	0.640	0.358	0.213
27539	-0.411	0.663	0.361	0.254
27540	-0.326	0.722	0.360	0.366
27545	-0.444	0.642	0.358	0.215
27560	-0.309	0.734	0.366	0.399
27562	-0.468	0.626	0.430	0.276
27571	-0.395	0.674	0.377	0.295
27587	-0.436	0.647	0.358	0.223
27591	-0.321	0.726	0.358	0.371
27592	-0.337	0.714	0.362	0.351
27596	-0.333	0.717	0.395	0.400
27597	-0.448	0.639	0.359	0.213
27601	-0.292	0.747	0.364	0.422
27603	-0.416	0.659	0.358	0.245
27604	-0.451	0.637	0.358	0.208
27605	-0.375	0.688	0.374	0.317
27606	-0.469	0.626	0.359	0.192
27607	-0.560	0.571	0.362	0.121
27608	-0.578	0.561	0.362	0.110
27609	-0.490	0.612	0.360	0.173

Table S6. Model 1, stage 1 (influence of water source and other variables on blood lead in early childhood)[a]

Variable	Coefficient (n=13,580)[b]	Odds Ratio	Standard Error	*p*
27610	-0.481	0.618	0.359	0.180
27612	-0.542	0.582	0.360	0.132
27613	-0.531	0.588	0.359	0.139
27614	-0.462	0.630	0.360	0.200
27615	-0.507	0.602	0.359	0.158
27616	-0.516	0.597	0.358	0.149
27617	-0.556	0.573	0.363	0.126

[a] Model R^2=0.13

[b] Coefficients from model of natural log of blood Pb regressed on all variables indicated.

Table S7. Model 1, stage 2, any complaint (influence of blood Pb on risk that a child will be reported for any juvenile complaint at age 14 or older)[a]

Variable	Coefficient (n=13,580)[b]	Odds Ratio	Standard Error	*p*
Blood Pb (natural log)	1.18	3.27	0.238	<0.001
Age indicator	1.12	3.05	0.261	<0.001
Male sex at birth (reference=female)	0.921	2.51	0.112	<0.001
Black race (reference=all other)	1.02	2.78	0.129	<0.001
Census block group median income (natural log) at age 14	-0.828	0.437	0.142	<0.001
Census block group % Black at age 14	0.178	1.19	0.277	0.521

[a] McFadden's R^2=0.15

[b] Coefficient from logistic regression model with any reported delinquency as dependent variable

Table S8. Model 1, stage 2, serious complaint (influence of blood Pb on risk that a child will be reported for a serious juvenile complaint at age 14 or older)[a]

Variable	Coefficient (n=13,580)[b]	Odds Ratio	Standard Error	*p*
Blood Pb (natural log)	1.48	4.39	0.324	<0.001
Age indicator	0.858	2.36	0.365	<0.05
Male sex at birth (reference=female)	1.20	3.33	0.174	<0.001
Black race (reference=all other)	1.12	3.05	0.194	<0.001
Census block group median income (natural log) at age 14	-0.843	0.431	0.196	<0.001
Census block group % Black at age 14	0.252	1.29	0.411	0.540

[a] McFadden's R^2=0.17

[b] Coefficient from logistic regression model with serious reported delinquency as dependent variable

Code and De-Identified Data Set

Model code and a partial data set (excluding information about the child at age 14) is available at the following web link: https://scholarworks.iu.edu/dspace/handle/2022/27027.

The full data set is available from the NC Education Research Data Center, through completion of the application for data use and the process described here: https://childandfamilypolicy.duke.edu/research/nc-education-data-center/.