WARWICK
ECONOMICS

Hidden hazards and Screening Policy : Predicting Undetected Lead Exposure in Illinois Using Machine Learning

Ali Abbasi, Ludovica Gazze & Bridget Pals

March 2022 No: 1398

Warwick Economics Research Papers

ISSN 2059-4283 (online)

ISSN 0083-7350 (print)

Hidden hazards and Screening Policy: Predicting Undetected Lead Exposure in Illinois Using Machine Learning*

Ali Abbasi
Department of Surgery, University of California San Francisco
Ali.Abbasi@ucsf.edu

Ludovica Gazze
Department of Economics, University of Warwick
Ludovica.Gazze@warwick.ac.uk

Bridget Pals
School of Law, New York University
bridget.pals@nyu.edu

January 28, 2022

Abstract
Lead exposure remains a significant threat to children's health despite decades of policies aimed at getting the lead out of homes and neighborhoods. Generally, lead hazards are identified through inspections triggered by high blood lead levels (BLLs) in children. Yet, it is unclear how best to screen children for lead exposure to balance the costs of screening and the potential benefits of early detection, treatment, and lead hazard removal. While some states require universal screening, others employ a targeted approach, but no regime achieves 100% compliance. We estimate the extent and geographic distribution of undetected lead poisoning in Illinois. We then compare the estimated detection rate of a universal screening program to the current targeted screening policy under different compliance levels. To do so, we link 2010-2016 Illinois lead test records to 2010-2014 birth records, demographics, and housing data. We train a random forest classifier that predicts the likelihood a child has a BLL above 5µg/dL. We estimate that 10,613 untested children had a BLL≥5µg/dL in addition to the 18,115 detected cases. Due to the unequal spatial distribution of lead hazards, 60% of these undetected cases should have been screened under the current policy, suggesting limited benefits from universal screening.

Keywords: Lead Poisoning, Environmental Health, Screening

* We are grateful to the Illinois Department of Public Health for providing the data used in this analysis, information on institutional background, and feedback on the findings, yet the conclusions, opinions, and recommendations in this paper are not necessarily theirs. The Joyce Foundation provided generous support. We are thankful to Stephen Billings, Francis DiTraglia, Andrew Oswald, Billy Pizer, David Slusky, and conference participants at APHA for helpful comments and suggestions. All errors are our own.

Introduction

A recent literature has emphasized the role of place in shaping children's opportunities growing up, however "we know little about the relative importance of the different mechanisms that are typically "bundled" together within a neighborhood", including pollution. (1) Lead is a neurotoxic heavy metal that was commonly used, for example in paint, gasoline, and plumbing. Because it does not decay, it still plagues neighborhoods throughout the United States, contaminating homes, soil, and water, and endangering human health. Childhood lead exposure is especially harmful; it is associated with lifelong developmental impacts, including decreased IQ and increased impulsivity and delinquency. (2–9) These burdens are disproportionately borne by communities of color and families of low socioeconomic status (10) potentially exacerbating existing inequalities. (11)

Lead paint was extensively used in the first half of the last century, until its ban for residential purposes in 1978 due to its recognized health hazards. The US Department of Housing Development estimates that lead paint still lingers in 5.5 million houses inhabited by small children nationwide, resulting in hazards in a fifth of homes with small children and constituting the major source of lead exposure today (12,13). Recognizing these risks, federal and state agencies continue to enact and fund policies to "get the lead out", including disclosure and abatement mandates of *known* lead hazards in homes. Yet, these policies appear to have failed to eliminate lead exposure and 500,000 young children are estimated to be still poisoned by lead each year in the US. (3)

While the reasons behind this policy failure remain unclear, this paper sheds light on one potential mechanism, namely imperfect information on the location of lead hazards. For example, a recent study uses the National Health and Nutrition Examination Survey to estimate that states detect and report to the CDC only 64% of the actual cases of BLL≥10μg/dL. (14) To analyze the role of hidden hazards and undetected lead poisoning, we implement a machine learning algorithm that identifies children who were likely exposed to lead but never tested. Our results highlight that the spatial distribution of lead exposure sources can be leveraged to target resources to uncover and remediate lead hazards. Until then, hidden lead hazards likely contribute to persistent patterns of spatial inequality.

Lead poisoning prevention programs in the United States follow a secondary prevention approach: blood lead screening identifies lead-exposed children who are then referred for case management including removing exposure sources. Which children are screened is thus crucial to identify lead hazards. While federal guidelines mandate that children on Medicaid are screened at ages one and two, guidelines for other children vary by state. Fourteen states and the District of Columbia currently mandate universal screening, that is testing all children, although their screening rates fall well short of 100%. (15) Other states have adopted the targeted screening approach recommended by the CDC, wherein testing is required only for children deemed at high risk for lead exposure, identified either through to socioeconomic and location information or through a self-assessment questionnaire. (16)

While targeted screening might better balance the costs of screening with the potential benefits of early detection and treatment, its efficacy hinges on the targeting tools, including self-

assessment questionnaires (17,18) and existing estimates of the distribution of exposure risks. However, these estimates might be biased precisely because targeted screening data cover a non-representative sample of children. (19) Moreover, the CDC targeting guidelines were last updated when the intervention threshold was 10μg/dL, (20) so at today's 5μg/dL reference level, or a proposed threshold of 3.5μg/dL, the benefits of targeted screening may diminish. (21)

We propose a new framework to estimate local childhood lead exposure prevalence in Illinois and use these estimates to compare detection rates under different screening policies. We use machine learning to predict the number of above-threshold BLLs missed under the current targeted screening policy and estimate how many of these additional above-threshold BLLs would be detected under different levels of compliance with a targeted vs. a universal screening policy. Currently, the Illinois Department of Public Health (IDPH) designates zip codes as high-risk based on housing age and the percentage of children living below 200% of the federal poverty line (Figure 1). Children must receive a BLL test if they reside in one of these high-risk zip codes, if they are on Medicaid, or if they screen positive on a risk assessment questionnaire. During most of our sample period (2010-2016), the intervention threshold in Illinois was 10μg/dL, but from 2015 local delegate agencies could lower the threshold based on their funding. In 2019, the intervention threshold was lowered to 5μg/dL for the entire state of Illinois.

Our model predicts the probability that each child born in Illinois between 2010 and 2014 had a BLL≥5μg/dL and ≥10μg/dL, regardless of whether they were screened for lead exposure. We train this model using lead testing records linked to geocoded birth records as well as other characteristics that are understood to contribute to lead exposure (housing age, proximity to

major roads, and industrial lead emissions). Using these predictions, we estimate the number of undetected above-threshold BLLs (both at BLL≥5μg/dL and ≥10μg/dL) under the status quo targeted screening strategy, as well as under a universal screening strategy. To account for imperfect compliance with universal screening, we repeat this estimation under various assumptions about screening compliance. By performing our analysis with two intervention thresholds of 10μg/dL and 5μg/dL, we evaluate how the effectiveness of lead screening policies changes as the intervention threshold decreases and different exposure sources assume different importance in explaining high blood lead levels. While we think that our preferred methodology best addresses issues of selection bias and rare outcomes, we also show that a simple logistic regression model could achieve reasonable accuracy and be used in a similar fashion. Thus, our methodology appears accessible to policy makers and stakeholders nationwide.

We report two main findings. First, we find evidence of significant non-detection: we estimate that over a third of cases of BLLs at or above 5μg/dL went undetected during our sample period. Second, undetected lead exposure cases appear to be disproportionally located in high-risk zip codes, potentially contributing to inequality in human capital and labor market outcomes. As a result, increasing screening rates in areas already targeted for screening would uncover more cases than extending universal screening at current compliance rates. Second, elevated BLLs at lower thresholds appear more geographically dispersed, suggesting that more subtle exposure sources driving low-level lead exposure might be less spatially clustered than main exposure sources at higher levels.

This paper contributes to a literature examining the benefits of universal screening. So far, studies have projected undetected cases by multiplying the number of children not tested under a targeted model with the rate of elevated test results. (22,23) This approach dramatically overestimates the benefits of universal screening because it assumes tested and untested children have similar rates of above-threshold BLLs. By contrast, we acknowledge 1) that children tested under a targeted screening scheme are, by construction, a higher risk group and 2) that compliance with screening guidelines is imperfect. (24,25) We also contribute to growing literature using machine learning tools to estimate the extent of lead exposure and innovate by using geocoded individual-level data. (26)

Data

We obtained birth records for all 807,694 children born in Illinois between 2010 and 2014 from IDPH. These birth records include the child's address, race, ethnicity, parental education level, parental age, and other demographic information. We also obtained records of all 1,105,168 lead tests performed in Illinois between 2010 and 2016 on children born between 2010 and 2014. Each lead test record contains the name of the child, the date of the blood draw, the type of blood test used (venous or capillary), the measured BLL, and the laboratory that processed the test. We use the highest BLL recorded for each child by age two to measure whether the child ever tested above the intervention threshold by that age. We use testing by age two both because the damage of lead exposure is thought to be more severe at lower ages, and to align with the federal screening guidelines for children on Medicaid, which specifically require two tests by age two. Because venous tests are more reliable than capillary tests, we consider each child's highest

venous test result. If a child had not received a venous test, we instead use the highest capillary test result.

Certain laboratories had minimum reporting limits, meaning all BLLs below a certain threshold were reported as the threshold limit (e.g. reporting BLL≤3μg/dL as 3μg/dL). We determine minimum reporting cutoffs for each laboratory/test type/year combination by manually reviewing BLL histograms. The BLL distribution is right-skewed, meaning an absence of tests below a certain value for a given laboratory likely indicates a minimum reporting limit. We estimate that 22,609 tests (2%) were performed by laboratories with a reporting limit ≥5μg/dL. We recode these to the mean BLL of children in the same zip and age cohort.

We link the lead testing and birth datasets using a custom fuzzy matching algorithm based on the Jaro-Winkler string distance of first name, last name, and date of birth, with manual determination of optimal cutoffs. (27) We successfully geocode birth addresses for 734,699 children. For each census block group, the American Communities Survey provides data on socioeconomic status, percent homeowners, and social vulnerability index. (28) We obtained data on housing age from the Zillow Transaction and Assessment Dataset, (29) and geocode these data for linkage to birth addresses. We also collected the Environmental Protection Agency's Toxic Release Inventory (TRI) data which detail industrial lead emissions by facility, (30) and the location of state and interstate highways from the Illinois Department of Transportation. (31) We then calculate the distance from lead-emitting facilities and roadways to each child's address.

Table 1 shows summary statistics for selected characteristics of children in our sample, stratified by whether a child was tested for lead exposure by age two. Tested children are more likely to be Black (20.4% vs. 14.1% $p<0.001$), Hispanic (28% vs. 16.2%, $p<0.001$), and have mothers without college education (43.9% vs. 27.9%, $p<0.001$). At birth, they are more likely to live in low-income census block groups (32.7% vs. 16.0%, $p<0.001$) and in pre-1930 housing (39.5% vs. 22% $p<0.001$).

Methodology

We use machine learning tools to predict the number of above-threshold BLLs amongst unscreened children based on observable characteristics in our data. Figure 2 summarizes the components of our prediction model. To begin, we generate 366 features characterizing demographics and exposure sources for each child. Because our goal is to build a model that could make accurate predictions on lead exposure among untested children, we train our model only on data that is available for both tested and untested children. Next, we use a Least Absolute Shrinkage and Selection Operator (LASSO) regression cross-validated at the minimum mean squared error to select 260 variables with a non-zero coefficient. We use these variables to train a random forest classifier that predicts the likelihood a child has a BLL≥5µg/dL and ≥10µg/dL. Weighted random forests avoid overfitting data and are particularly useful for datasets with a rare outcome of interest, as is an above-threshold BLL. (32) We train our model on 90% of the records of tested children, randomly withholding 10% for validation.

Our prediction problem presents two main challenges: selective labels and rare events. First, we can only train our model on tested children, and these children are observably different from untested children (see Table 1). To mitigate selection bias, we re-weight each observation *i* in the training sample by the inverse probability that child *i* was screened, P^{i}_{screen}. To estimate P^{i}_{screen}, we use a LASSO logistic regression and random forest, using the observable characteristics discussed above, as well as additional predictors of screening such as distance from a testing provider.

Second, above-threshold BLLs are a comparatively rare event, occurring in less than five percent of the study population. This means an algorithm could achieve 95% accuracy by simply classifying *all* samples as below-threshold. To improve prediction accuracy for above-threshold BLLs, we oversample above-threshold BLLs in the training data so that the model would encounter a roughly equal number of below- and above-threshold BLLs. (32) We re-weight above-threshold BLL observations by the inverse probability of above-threshold BLLs, $P_{BLL>Threshold}$. The final weight of each child *i* in the training data is $1/ P^{i}_{screen}$ if the highest BLL was below-threshold and $1/(P^{i}_{screen} * P_{BLL\geq Threshold})$ if the highest BLL was above-threshold.

With these weights, we use the R ranger package to train the random forests and tuneRanger to optimize the random forest parameters. (33) We validate the model's performance using the 10% sample of randomly withheld testing data. Among the 1% of children labelled as highest risk, 39.7% and 25.6% had a BLL≥5µg/dL and ≥10µg/dL respectively, compared to 3.9% and 0.6% of all children in the sample (Figure A.1). Thus, amongst the highest risk children the model performs ten times and forty-two times better than random chance at predicting BLL≥5µg/dL

and BLL≥10μg/dL, respectively. We report variable importance using Gini importance. The ten highest Gini importance variables for prediction of BLL≥5μg/dL include the number of above-threshold BLLs near the child's birth address and socioeconomic status in the child's birth block group (Table A.1).

Finally, we compare our random forest model to a LASSO penalized linear regression model that predicts BLLs, as well as to a logistic regression.[1] Both the linear and logistic models are estimated on the unweighted training sample. Table A.2 shows estimates from the logistic model for selected predictors that appear to significantly correlate with the probability that a child had BLL≥5μg/dL. Most relationships have the expected sign: younger and less educated mothers are more likely to have children with a BLL≥5μg/dL, although the same is true for married mothers, which is unexpected. Children in bigger families, older homes, high risk neighborhoods, and homes and neighborhoods with past cases of high BLLs are more likely to have high blood lead levels. However, in buildings with more than one instance of previous cases of high BLLs, this probability decreases, potentially due to remediations. It is noteworthy that coefficients on neighborhood socioeconomic status variables have an unexpected sign, likely due to predictors being highly correlated. Proximity to major roads and industrial establishments emitting lead does not appear to be significantly correlated with above-threshold BLLs.

We examine the performance of these models using the area under the curve (AUC) on the receiver operating characteristic curve (ROC), which plots the true positive rate (sensitivity) vs.

[1] The logistic regression excludes the 103 county indicators included in the other models for computational reasons.

false positive rate (1-specificity) of a prediction model for all possible cutoff values of the predictor (Figure A.2). The random forest outperforms the linear regression model in predicting both BLL≥10μg/dL and ≥5μg/dL (Figure A.2).[2] The random forest achieves AUCs of 0.774 and 0.905 for predicting BLLs ≥5μg/dL and 10μg/dL on the testing sample, respectively. Interestingly, the simpler logistic model slightly outperforms the random forest model in predicting BLL≥5μg/dL, but does worse in predicting BLL≥10μg/dL, achieving AUCs of 0.778 and 0.896, respectively.

To estimate a child's actual exposure risk, we divide the sample into 50 risk groups based on the value of their random forest predictor, separately for BLL≥10μg/dL and ≥5μg/dL. We compute an untested child's exposure risk as the average risk among tested children within the same risk group.

To account for imperfect compliance, we simulate the number of above-threshold BLLs detected under each model for four levels of compliance with screening guidelines, that is the 50th, 75th, 90th, and 100th percentile of screening rates in current high-risk zip codes, corresponding to screening rates of 61.1%, 71.4%, 81.3% and 100% (Figure A.3). In each zip code with population P and T tested children, we randomly select U of the (P-T) untested children until the total screening rate $(U+T)/P$ is equal to the desired screening rate. For each of these U children, we simulate whether the child had an above-threshold BLL based on our calibrated probability. We calculate the total number of above-threshold BLLs in each zip code by adding the number

[2] For this exercise, we use the rank order of predicted BLLs from the linear model.

of above-threshold BLLs detected among *T* tested children to the number of above-threshold BLLs simulated among the *U* untested children. We repeat the sampling process 1,000 times and reported mean values and 95% confidence intervals.

Results

In our sample, 18,115 tested children have a BLL≥5µg/dL and 3,292 tested children have a BLL≥10µg/dL. We estimate substantial underdetecton of lead exposure: among children born between 2010-2014, current testing practices detected 63% of BLL≥5µg/dL and 70% of BLL≥10µg/dL. Indeed, our model predicts an additional 10,613 (95%CI 10,423-10,804) of the 356,432 untested children had BLL≥5µg/dL (Table 2). We also predict an additional 1,387 (95%CI 1,319 to 1,455) of the 356,432 untested children had BLL≥10µg/dL.

While the number of predicted above-threshold BLLs decreased significantly over time, the detection rate stayed relatively stable. Figure 3 plots the model-predicted number of total above-threshold BLLs, by year, by screening compliance, and by BLL intervention threshold (5 and 10µg/dL). In the 2010 birth cohort, there were 7,348 children with a BLL≥5µg/dL, compared to 4,529 children in the 2014 cohort – a 38.4% decline (Figure 3a). However, the percent of all BLLs that were detected rose only from 60.5% in 2010 to 63.9% in 2014.

To investigate where children with undetected BLLs are, Figure 4 plots the rate of detected and predicted undetected BLLs≥5µg/dL in each zip code, highlighting a positive correlation between detected and undetected cases. Undetected lead poisoning cases appear to be concentrated in

areas already identified as high risk and therefore these children should have been tested under Illinois' existing screening policy. Indeed, a comparison with Figure 1 shows a disproportionate rate of undetected above-threshold BLLs in high-risk zip codes. Table A.3 further illustrates this point by showing the model predictions stratified by zip code risk status. While there are fewer untested children in high-risk zip codes (119,077 vs. 237,355), 60% of untested children with predicted BLL≥5μg/dL lived in high-risk zip codes (6,375) rather than low-risk ones (4,238). Our model predicted that 68% of children with undetected BLL≥10μg/dL were in high-risk zip codes (948) rather than low-risk zip codes (440).

This unequal distribution of lead hazards suggests that targeted screening might have its merits. Figure 5 plots the geographic dispersion of predicted above-threshold BLL by calculating the cumulative number of above-threshold BLL located in each top percentile (x-axis) of zip codes based on the probability of above-threshold BLL. Currently, Illinois designates 42% of zip codes as high-risk. These zip codes include 66.3% and 71.5% of all predicted cases of BLL≥5μg/dL and BLL≥10μg/dL. However, we find evidence of spatial concentration of lead hazards translating into unequal likelihood of lead exposure even with neighborhoods already deemed as high risk. According to our predictions, the riskiest 42% of zip codes in Illinois were home to almost 95% of all children with BLL≥10μg/dL. Thus, by adjusting the definition of high-risk zip codes, states may be able to detect a higher number of above-threshold BLLs without increasing the number of high-risk zip codes.

Next, we investigate the role of compliance with screening guidelines in leaving cases of above-threshold BLLs undetected. If a universal screening regime had been in place and all zip codes

achieved the same screening compliance as the median high-risk zip code under Illinois' existing regime, an additional 89,761 children would have been tested, which would have resulted in detecting only an additional 1,537 BLL≥5µg/dL (95%CI 1,463-1,612) and an additional 168 BLL≥10µg/dL (95%CI 143-193). Raising compliance to the 75th percentile would have required 150,237 additional children to be tested and would have found another 3,236 BLL≥5µg/dL (95%CI 3,129-3,343) and an additional 389 BLL≥10µg/dL (95%CI 351-426). Compliance with universal screening at the 90th percentile of current high risk zip codes would have resulted in an additional 219,138 tests and detected another 5,572 BLL≥5µg/dL (95%CI 5,432-5,712) and 704 BLL≥10µg/dL (95%CI 654-753). We estimate that 76% of children with undetected BLLs≥10µg/dL and 61% of children with undetected BLLs≥5µg/dL resided in high-risk zip codes.

These estimates suggest that universal screening is not necessarily a panacea to fix under-detection. Under realistic assumptions about compliance with universal screening, only a fraction of these additional above-threshold BLLs would have been detected. It is possible that a universal screening policy could increase compliance because it is easier to communicate but such gains in compliance are far from assured. All zip codes in Chicago are high-risk, implying a de-facto universal screening policy. Yet, average screening rates in Chicago were 62.6%, compared to 62.9% in high-risk zip codes outside of Chicago.

Our results highlight one challenge in identifying and addressing lead hazards going forward. First, as the exposure threshold decreases, above-threshold BLLs become more dispersed throughout the state (Figure 5). As illustrated in Table 2, 50% and 90% of all predicted

BLL≥5µg/dL cases resided in 32.2% and 70.8% of all zip codes, while 50% and 90% of all BLL≥10µg/dL cases resided in 16.9% and 34.7% of all zip codes. Perhaps as a result, a higher share of BLL≥5µg/dL cases than BLL≥10µg/dL cases go undetected. During the study period, there were approximately 6.5 times as many cases of BLL≥5µg/dL as BLL≥10µg/dL, but we estimate 10 times as many *undetected* cases of BLL≥5µg/dL as BLL≥10µg/dL. In related work, we have also shown that the relative importance of exposure sources shifts with decreasing intervention thresholds, which may make it more difficult to identify cases of above-threshold BLLs by relying on proxies for lead exposure. (21) The increased dispersion of cases that exceed lower intervention thresholds reduces the benefits of targeted screening.

Limitations

To predict above-threshold BLLs among all children, we are limited to predictors of lead exposure at the birth address, because addresses at the time of test are only available for tested children. We also do not have access to reliable data indicating whether or not a child was on Medicaid, a potentially important determinant of testing. While we validate the predictive accuracy of these variables on withheld data, data reflecting current residences could further improve predictive power. Our analysis also lacks data on certain pathways for lead exposure, such as lead in drinking water, toys, or parental occupational exposure. However, housing vintage likely partially accounts for the effects of lead in water because the use of lead pipes and service lines follows historical patterns. (34) Additionally, the missing exposure sources are understood to represent only a small part of total lead exposure. (10) Finally, while we mitigate

selection biases by re-weighting our prediction sample, there are observable differences in risk between tested and untested children.

Policy Implications

Figure 3 shows that the share of non-detected above-threshold BLLs remained constant throughout our sample period, although lead exposure in Illinois and nationwide has decreased. (20) This finding suggests that under-detection will remain a substantial issue even as the absolute number of cases continues to fall. As case management has been found to reduce the damages of lead exposure, under-detection could be a significant factor in hindering these children's ability to develop. (35) So, is universal screening warranted?

In this section, we perform a back-of-the-envelope calculation to compare the cost of detecting an additional child with above-threshold BLL under a universal and targeted screening program focusing on current high-risk zip codes in Illinois. The results crucially hinge on our main finding that lead exposure appears concentrated in neighborhoods already deemed as high risk, thus suggesting that a targeted approach might be more cost-effective. However, as the intervention threshold decreases more children would be eligible for interventions (and thus benefit from screening) outside areas currently considered as high-risk, increasing the value of a universal screening program.

The costs and benefits of a screening program are hard to quantify. Costs include laboratory and material costs, opportunity costs of time for parents, e.g., travel costs to the doctor's office (36)

and health care service providers, and non-monetary costs (e.g., pain if venous blood sample). The price tag for private tests ranges up to $43 in Illinois[3], while other indirect costs are harder to measure. We estimate that under universal screening with perfect compliance, an additional 1,387 cases of BLL≥10µg/dL would have been detected among the 356,432 untested children born in Illinois between 2010 and 2014. Thus, the nominal cost of tests (using $43 as the full cost) per case of BLL≥10µg/dL detected is $11,050. A program increasing screening to 100% in high-risk zip codes only would have detected 948 additional cases of BLL≥10µg/dL among the 119,077 unscreened children born in those zip codes, at a cost of $5,401, that is less than a half of the cost of a universal screening program. The cost per BLL≥5µg/dL detected would be $1,444 under a universal screening program and $803 under a targeted program achieving 100% compliance, allowing to detect 10,613 and 6,375 additional cases, respectively.

As a benchmark, we can consider that the main benefits of screening accrue to children with above-threshold BLLs who receive beneficial interventions. These interventions have been found to improve school performance and reduce antisocial behavior, for a value of $9,666 for each child with a BLL≥10µg/dL. (35)

Conclusion

We estimate the extent and geographic distribution of undetected lead poisoning in Illinois using administrative data and machine learning tools. We leverage these estimates to compare how

[3] https://www.luc.edu/media/lucedu/hhhci/pdf/leadsafeil/LeadSafeILDirectory061_.pdf, accessed in November 2021)

many above-threshold BLLs are missed under the *status quo* targeted screening and a simulated universal screening regime with different levels of compliance, for intervention thresholds of 5μg/dL and 10μg/dL. We find that current testing practices failed to detect 37% of BLL≥5μg/dL and 30% of BLL≥10μg/dL. Moreover, 60% of children with undetected BLL≥5μg/dL in Illinois lived in zip codes where every child should already be tested under current Illinois' testing guidelines. These are neighborhoods with old housing and low socioeconomic status, suggesting that undetected lead poisoning might exacerbate existing patterns of inequality.

The spatial distribution of lead hazards implies that states may see the largest gains in above-threshold BLL detection from improving compliance with existing screening policies, rather than expanding to a universal screening regime as currently advocated by many. How to increase screening rates remains an open question, however. Travel cost and inconvenient access to health care providers appear to be one barrier, together with providers' idiosyncratic lower propensity to refer children for lead screening. (36) We caveat our analysis noting that under a lower exposure threshold, the benefits of targeted screening may be reduced because above-threshold BLLs become more geographically disperse as the threshold is lowered.

Finally, we demonstrate how machine learning can improve targeted screening by leveraging detailed demographic and exposure data and providing a more accurate estimate of each child's risk of an above-threshold BLL. This risk estimate could be used to categorize zip codes as high-risk in a targeted screening program. While local health departments have used similar approaches to target screening and educate providers and patients about their risk, (37) our work represents the first such model at a state-level. Statewide models can harness economies of scale,

using larger datasets to improve prediction accuracy, while local health departments may lack the resources for geospatial modeling. Implementing risk stratification tools into electronic medical records could also help healthcare providers ensure that the highest risk children are tested. This avenue offers promise; Figure 1 shows that in our model, the highest risk 1% of children have a probability of above-threshold BLL≥5µg/dL of 39.7%, more than 10 times that of an average Illinois child, opening the door for child-level targeted screening. This approach can be adapted for other states based on the available data sources. The resulting predictions can be used to inform lead testing policy, evaluate the effects of changing intervention thresholds, and identify the children at highest risk for lead exposure.

References

1. Chyn E, Katz LF. Neighborhoods Matter: Assessing the Evidence for Place Effects. J of Econ Perspectives 2021;35(4):197-222.
2. Aizer A, Currie J. Lead and Juvenile Delinquency: New Evidence from Linked Birth, School and Juvenile Detention Records. Review of Economics and Statistics. 2019;101(4): 575–87.
3. Aizer A, Currie J, Simon P, Vivier P. Do Low Levels of Blood Lead Reduce Children's Future Test Scores? American Economic Journal: Applied Economics. 2018;10(1):307–41.
4. Feigenbaum JJ, Muller C. Lead Exposure and Violent Crime in the Early Twentieth Century. Explorations in Economic History. 2016;62:51–86.
5. Reyes JW. The Social Costs of Lead in Lead: The Global Poison – Humans, Animals, and the Environment. 2014;1–4.
6. Reyes JW. Lead Exposure and Behavior: Effects on Antisocial and Risky Behavior among Children and Adolescents. Economic Inquiry. 2015;*53*(3):1580–1605.
7. Bellinger DC, Stiles KM, Needleman HL. Low-Level Lead Exposure , Intelligence and Academic Achievement: A Long-term Follow-up Study. Pediatrics. 1992;90(6):855–61.
8. Winter AS, Sampson RJ. From Lead Exposure in Early Childhood to Adolescent Health: A Chicago Birth Cohort. Am J Public Health. 2017;107(9):1496–501.
9. Grönqvist H, Nilsson JP, Robling P-O. Understanding How Low Levels of Early Lead Exposure Affect Children's Life Trajectories. J Polit Econ. 2020;128(9):3376–433.
10. Zartarian V, Xue J, Tornero-Velez R, Brown J. Children's Lead Exposure: A Multimedia Modeling Analysis to Guide Public Health Decision-Making. Environ Health Perspect.

2017;125(9):097009.

11. Sampson RJ, Winter AS. Toxic Inequality in Chicago Neighborhoods. Du Bois Rev. 2016;1995–2013.
12. U.S. Department of Housing and Urban Development. American healthy homes survey lead and arsenic findings. 2011.
13. Dewalt FG, Cox DC, O'Haver R, Salatino B, Holmes D, Ashley PJ, Pinzer EA, Friedman W, Marker D, Viet SM, Fraser A. Prevalence of lead hazards and soil arsenic in U.S. housing. Journal of Environmental Health. 2015;78(5): 22–52.
14. Roberts EM, Madrigal D, Valle J, King G, Kite L. Assessing child lead poisoning case ascertainment in the US, 1999-2010. Pediatrics. 2017;139(5).
15. Michel JJ, Erinoff E, Tsou AY. More Guidelines than states: variations in U.S. lead screening and management guidance and impacts on shareable CDS development. BMC Public Health. 2020;20.
16. Centers for Disease Control and Prevention. Screening Young Children for Lead Poisoning: Guidance for State and Local Public Health Officials. Washington, DC; 1997.
17. Dyal B. Are Lead Risk Questionnaires Adequate Predictors of Blood Lead Levels in Children? Public Health Nurs. 2012;29(1):3–10.
18. Binns HJ, LeBailly SA, Fingar AR, Saunders S. Evaluation of risk assessment questions used to target blood lead screening in Illinois. Pediatrics. 1999;103(1):100–6.
19. Manheimer EW, Silbergeld EK. Critique of CDC's retreat from recommending universal lead screening for children. Public Health Rep. 1998;113(1):38–46.
20. Tsoi M-F, Cheung C-L, Cheung TT, Cheung BMY. Continual Decrease in Blood Lead

Level in Americans: United States National Health Nutrition and Examination Survey 1999-2014. Am J Med. 2016;129(11):1213–8.

21. Abbasi A, Pals B, Gazze L. Policy Changes and Child Blood Lead Levels by Age 2 Years for Children Born in Illinois, 2001–2014. Am J Public Health. 2020.

22. Maryland Department of Health and Mental Hygiene. Maryland targeting plan for children areas at risk for childhood lead poisoning. Baltimore; 2015.

23. McMenamin SB, Hiller SP, Shigekawa E, Melander T, Shimkhada R. Universal Lead Screening Requirement: A California Case Study. Am J Public Health. 2018;108(3):355–7.

24. Einav L, Finkelstein A, Oostrom T, Ostriker A, Williams H. Screening and selection: The case of mammograms. Am Econ Rev. 2020 Dec;110(12):3836-70.

25. Kim H, Lee S. When public health intervention is not successful: Cost sharing, crowd-out, and selection in Korea's National Cancer Screening Program. J Health Econ. 2017;53:100–16.

26. Lobo, GP, Kalyan B, Gadgil AJ. Predicting childhood lead exposure at an aggregated level using machine learning. International Journal of Hygiene and Environmental Health. 2021;238:113862.

27. Winkler WE. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. 1990.

28. U.S. Census Bureau's American Community Survey Office. B19013: Median household income in the past 12 months. 2015 American Community Survey. 2015.

29. Zillow Group Inc. Zillow Transaction and Assessor Dataset. 2017.

30. US Environmental Protection Agency. Toxics Release Inventory (TRI) Program.

31. Illinois Department of Transportation. Highway System.

32. Chen C, Liaw A, Breiman L. Using Random Forest to Learn Imbalanced Data | Department of Statistics. Berkeley, CA; 2004.

33. Probst P, Wright MN, Boulesteix AL. Hyperparameters and tuning strategies for random forest. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019.

34. Rabin R. The lead industry and lead water pipes "A Modest Campaign." Am J Public Health. 2008;98(9):1584–92.

35. Billings SB, Schnepel KT. Life after Lead: Effects of Early Interventions for Children Exposed to Lead. American Economic Journal: Applied Microeconomics. 2018;10(3): 315–44.

36. Gazze L. Hassles and Environmental Health Screenings: Evidence from Lead Tests in Illinois. Journal of Human Resources. Forthcoming.

37. Potash E, Brew J, Loewi A, Majumdar S, Reece A, Walsh J, Rozier E, Jorgenson E, Mansour R, Ghani R. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. KDD '15- Proc 21th ACM SIGKDD Int Conf Knowl Discov Data Min. 2015;2039–47.

Figures

Figure 1: High-Risk Zip Codes in Illinois (2006-Present Designation)

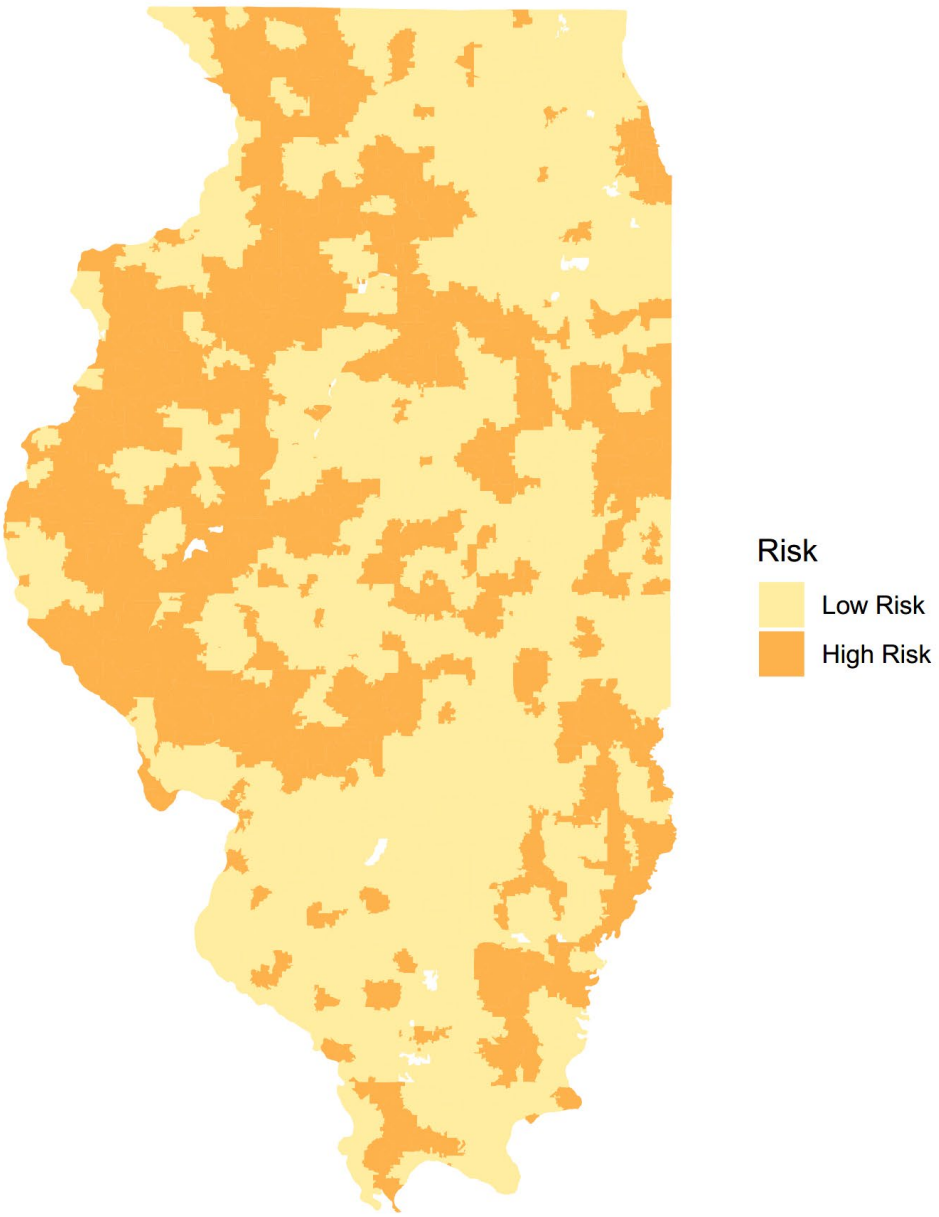
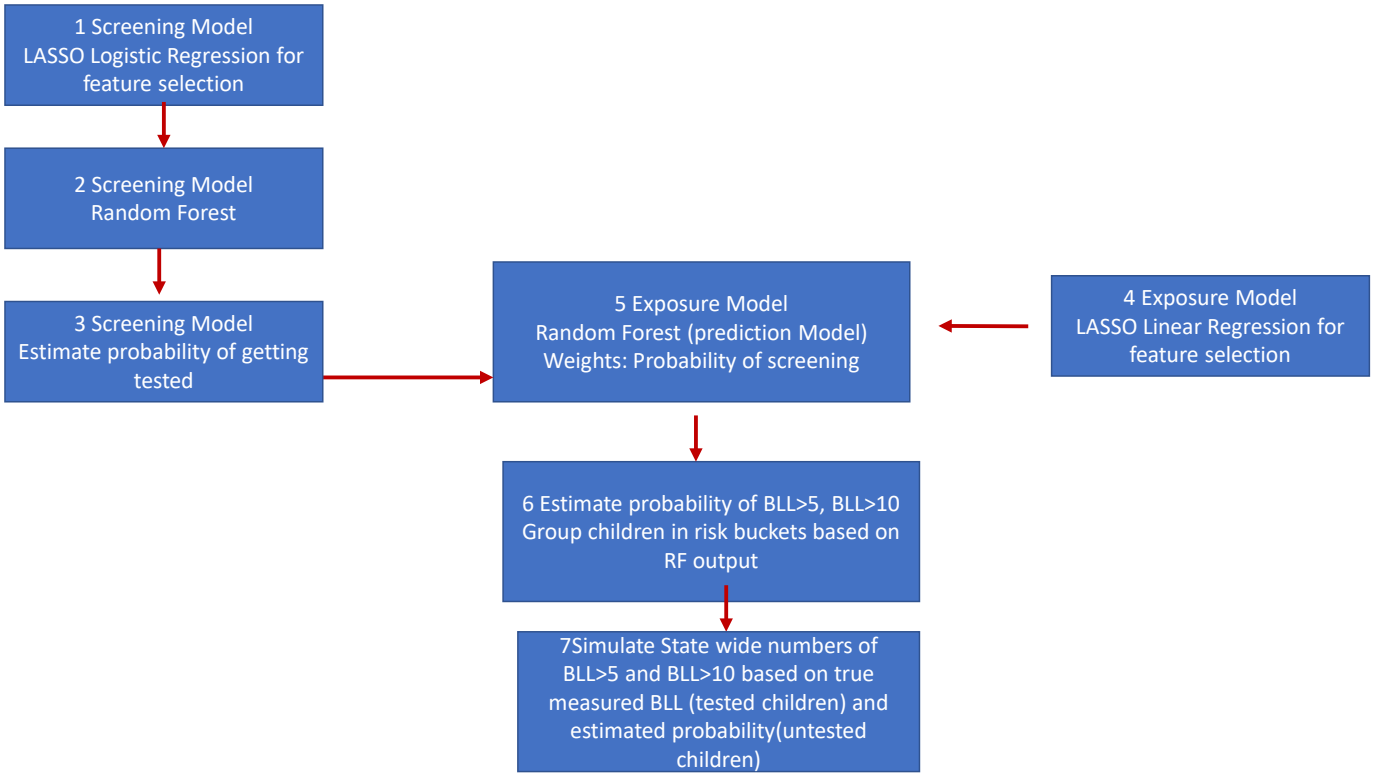
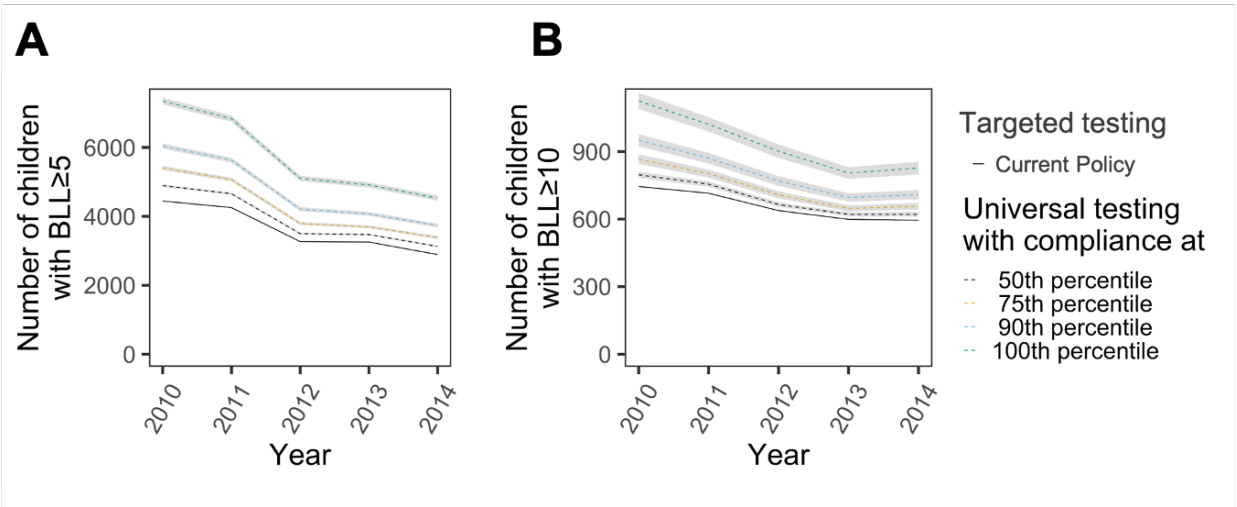


Figure 2: Predictive Model Architecture



Notes: The figure illustrates the steps that our predictive model entails.

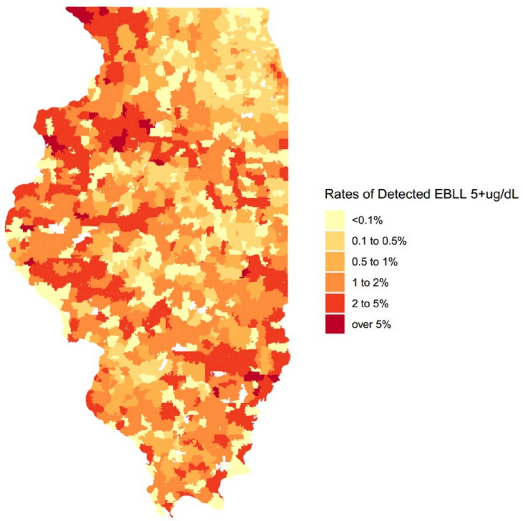
Figure 3: Number of Children with Detected BLLs≥5µg/dL (Panel A) and BLLs≥10µg/dL (Panel B) under Current and Counterfactual Screening Policies



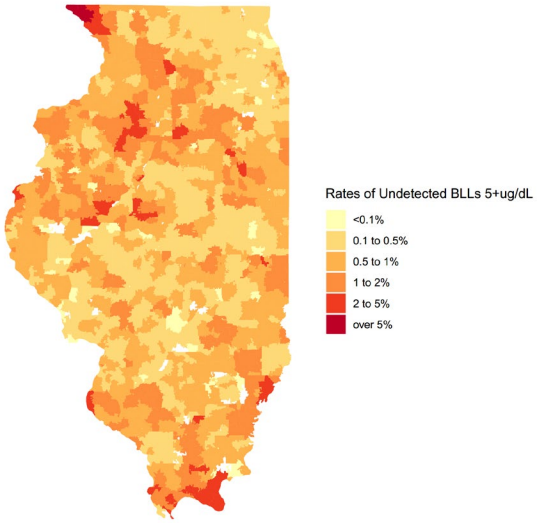
Notes: The figures plot the number of children with detected BLLs≥5µg/dL (Panel A) and BLLs≥10µg/dL (Panel B) under the current targeted screening policy and realized compliance (solid line) and counterfactual universal screening policies with compliance rates currently achieved by the zip code at the 50th, 75th, 90th, and top percentile of the screening rate distribution. Grey bands represent 95% confidence intervals.

Figure 4: Rates of Detected (Panel A) and Predicted Undetected (Panel B) BLLs≥5µg/dL by Zip Code

Panel A

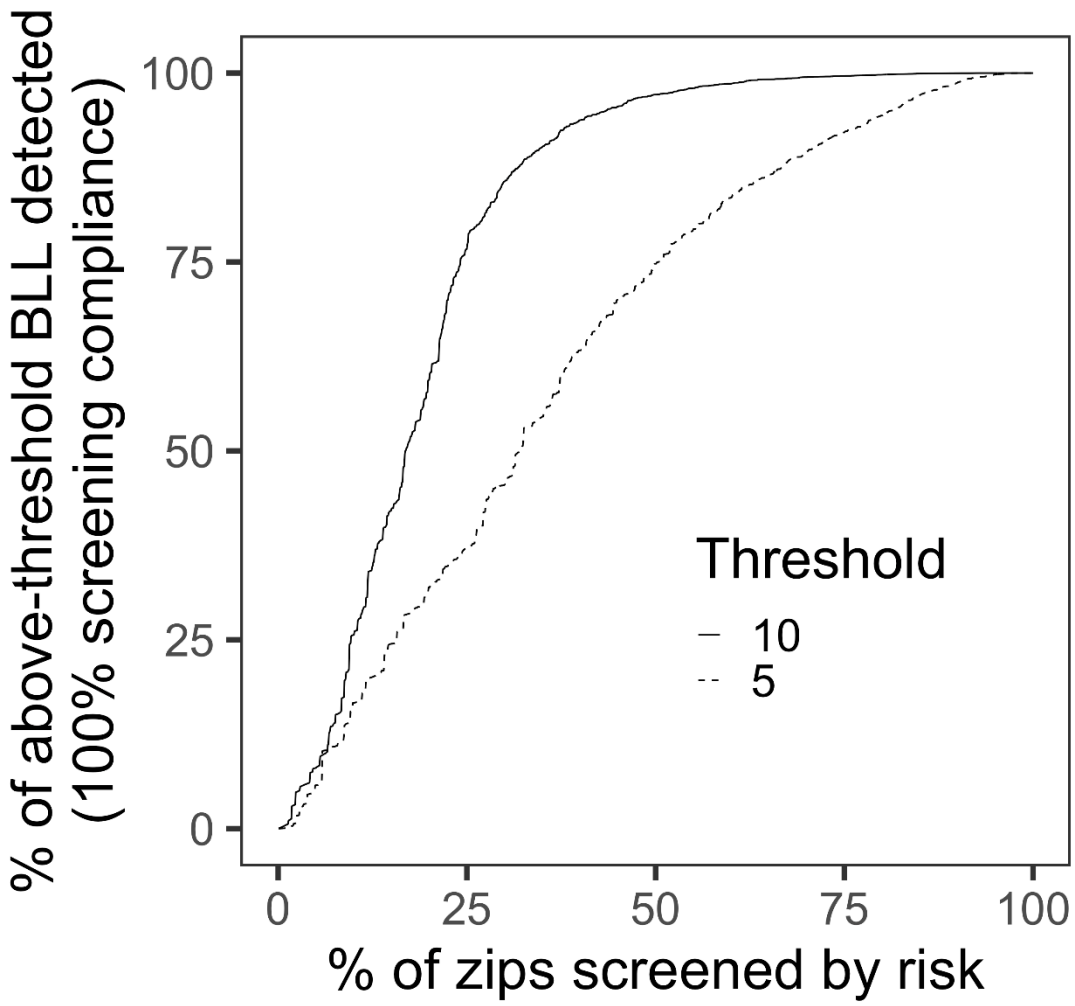


Panel B



Notes: The figures plot the number of children with detected BLLs≥5µg/dL (Panel A) and undetected BLLs≥5µg/dL (Panel B) over total children for birth cohorts 2010-2014.

Figure 5: Cumulative Share of Above-Threshold BLLs in Zip Codes Ranked by Number of Children with Above-Threshold BLLs



Notes: The figures plot the number of children with detected BLLs≥5µg/dL (Panel A) and BLLs≥10µg/dL (Panel B) under the current targeted screening policy and realized compliance (solid line) and counterfactual universal screening policies with compliance rates currently achieved by the zip code at the 50th, 75th, 90th, and top percentile of the screening rate distribution. Grey bands represent 95% confidence intervals.

Tables

Table 1
Baseline Characteristics of all children born in Illinois 2010-2014, stratified by whether they were born in a high risk zip code and whether they were tested for lead exposure by age 2.

	Low Risk		High Risk	
Screened	**No**	**Yes**	**No**	**Yes**
N	237355	178010	119077	200257
Black (%)	20035 (8.5)	22321 (12.6)	29813 (25.2)	54313 (27.3)
Hispanic (%)	30278 (12.8)	39893 (22.5)	27527 (23.2)	66171 (33.2)
Mother < 20 years at birth (%)	9322 (3.9)	14482 (8.1)	9872 (8.3)	20994 (10.5)
Mother Unmarried at birth (%)	57292 (24.1)	75636 (42.5)	52315 (44.0)	107128 (53.5)
Mother with no college education (%)	54050 (23.0)	69863 (39.6)	45468 (38.8)	96205 (48.7)
Median Income in Census Block (sd)	70604 (22215)	62881 (20746)	49657 (18002)	46182 (16351)
Any TRI release within 250 m (%)	347 (0.1)	321 (0.2)	719 (0.6)	1127 (0.6)
Birth address built before 1930 (%)	12474 (9.4)	13857 (14.2)	49119 (56.4)	90691 (63.4)
Birth address built before 1980 (%)	73550 (55.6)	71078 (73.0)	73188 (84.0)	126489 (88.4)
Birth zip code in Chicago (%)	55 (0.0)	62 (0.0)	73037 (61.3)	122059 (61.0)
Previous BLL≥5µg/dL at birth address (%)	2164 (0.9)	3321 (1.9)	14011 (11.8)	29314 (14.6)
Previous BLL≥10µg/dL at birth address (%)	801 (0.3)	1243 (0.7)	6176 (5.2)	13520 (6.8)
Highest BLL ≥ 5µg/dL (%)		5433 (3.1)		12682 (6.3)
Highest BLL ≥ 10µg/dL (%)		896 (0.5)		2396 (1.2)

Notes: SD Standard Deviation; BLL Blood Lead Level

Table 2

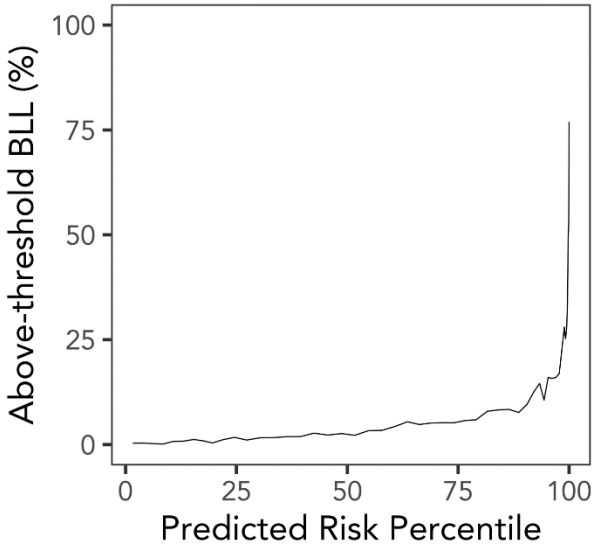
Caption: Simulated effect of universal screening at different rates of compliance with universal screening policy and intervention thresholds of 5µg/dL and 10µg/dL amongst children born in Illinois 2010-2014. Target screening rates were chosen to coincide with the 50th, 75th, 90th, and 100th percentile, of current high-risk zip codes in Illinois where all children should be tested, corresponding to screening rates of 61%, 71%,81%, 100%. Mean and 95% confidence interval are the results of 1000 simulations of testing additional children based on the probability of above-threshold BLLs derived from a calibrated random forest predictor.

Target screening rate (percentile)	**Intervention Threshold (µg/dL)**	**Number of children**	**Screened Children**	**Actual above-threshold BLLs**	**Additional Children Screened**	**Predicted additional above-threshold BLLs (Mean (95% CI))**
61% (50th)	5	734,699	378,267	18,115	89,761	1,537 (1,463, 1,612)
71% (75th)	5				150,237	3,236 (3,129, 3,343)
81% (90th)	5				219,138	5,572 (5,432, 5,712)
100% (100th)	5				356,432	10,613 (10,423, 10,804)
61% (50th)	10			3,292	89,761	168 (143, 193)
71% (75th)	10				150,237	389 (351, 426)
81% (90th)	10				219,138	704 (654, 753)
100% (100th)	10				356,432	1,387 (1,319, 1,455)

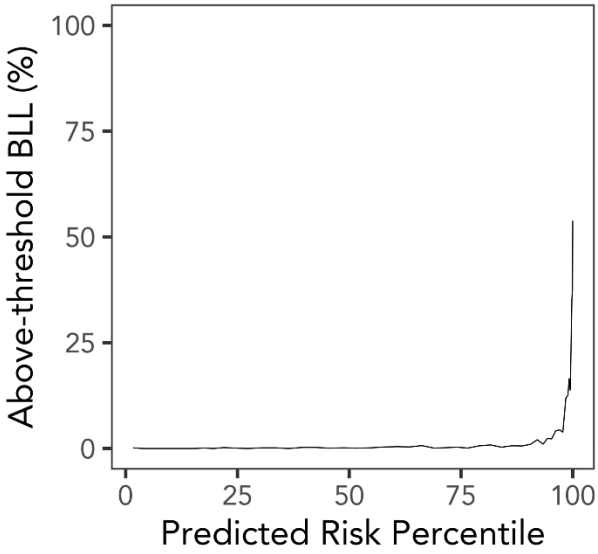
Supplementary Tables and Figures

Figure A.1: Performance of random forest predicting risk of above-threshold BLL.

Panel A: Performance for BLL≥5µg/dL



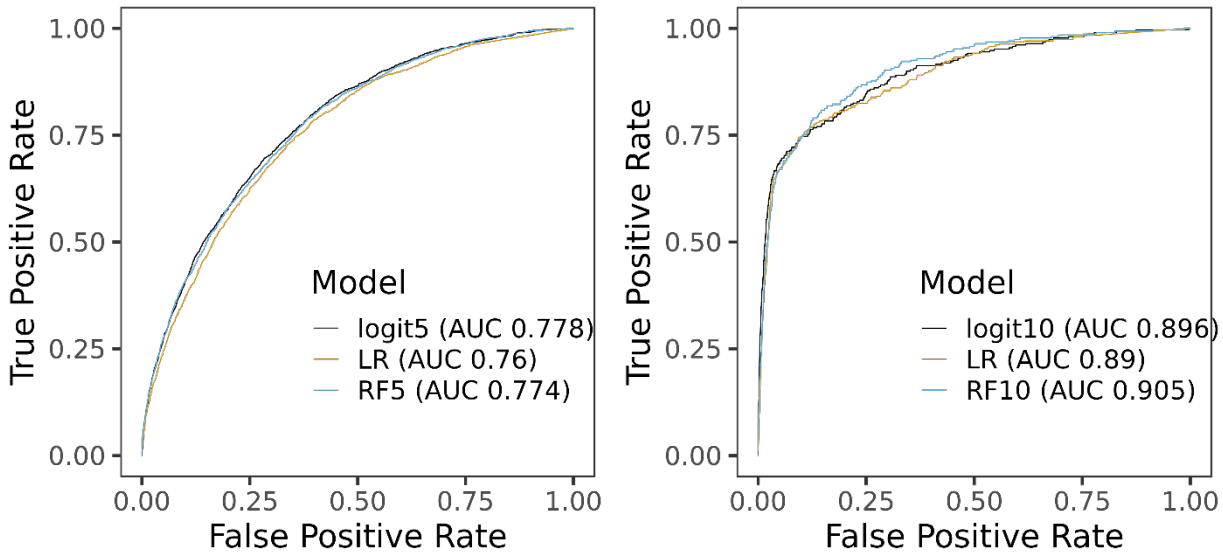
Panel B: Performance for BLL≥10µg/dL



Notes: Children in our sample were stratified into 50 risk groups. This figure shows the percentage of children in each group with an above-threshold BLL as confirmed by the testing data that was withheld from training. Panel A plots results for the threshold of 5µg/dL while Panel B for 10µg/dL.

Figure A.2: Model performance on Receiver Operating Characteristic Curve across Models

Panel A: Model performance for BLL≥5µg/dL Panel B: Model performance for BLL≥10mg



Notes: The figure plots the Receiver Operating Characteristic (ROC) Curve across the logistic (logit), linear regression (LR), and random forest (RF) models trained to predict BLL≥5µg/dL (Panel A) and BLL≥10µg/dL (Panel B). The ROC curve illustrates how the true (y axis) and false (x axis) positive rate for each model change with changes in the threshold of predicted scores used to classify observations as above-threshold BLLs. The figures also report the area under the ROC curve (AUC) for each model, a summary measure of goodness of fit.

Figure A.3: Histogram of screening rates in high-risk zip codes in Illinois 2010-2014

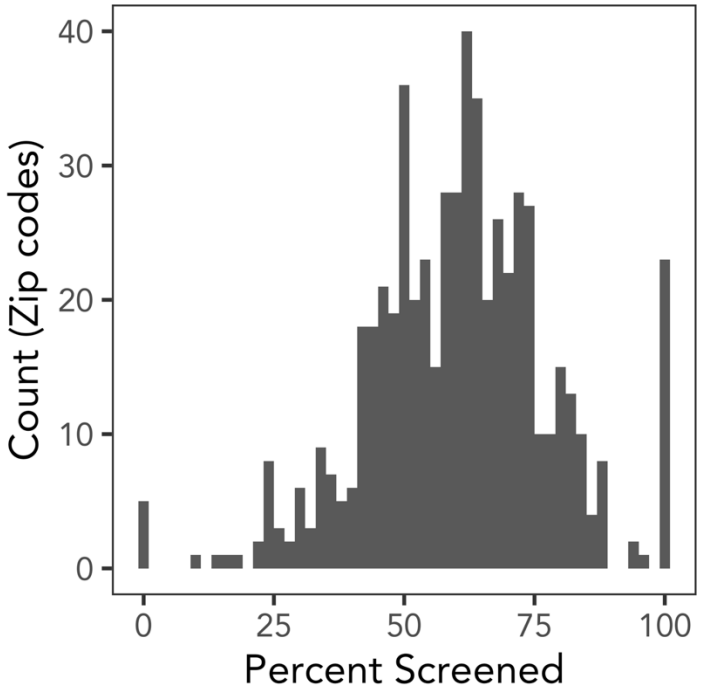
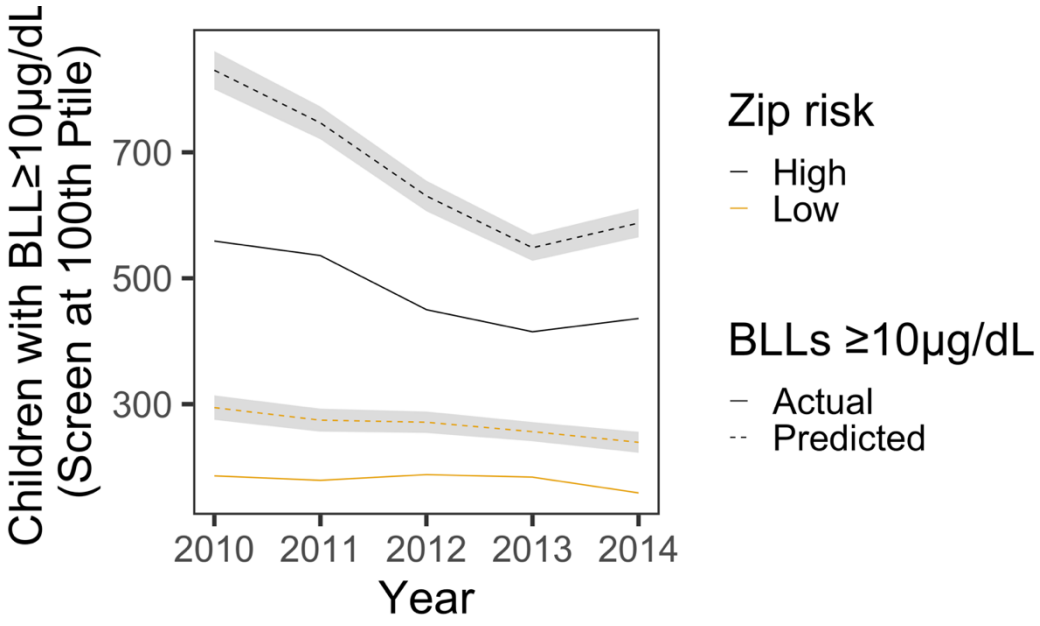


Figure A.4: Above-threshold BLLs detected and predicted, by Zip Code Risk



Notes: The figure plots the number of BLL≥10µg/dL actually detected (solid line) and predicted (dotted lines) in Illinois 2010-2014 stratified by risk status of zip code. Grey ribbons denote 95% confidence intervals.

Table A.1: Top ten most important variables used in prediction of BLL≥5µg/dL

Variable	LASSO Coefficient	Importance 5µg/dL
Percent of census tract BLL≥10	2.90	2392.67
Child with BLL≥5 within 15m and 1 year of child	1.90	512.08
BLL≥5 within 500m	0.04	413.35
Median HHD Income, block group	0.00	304.86
Median Home Value ($)	0.00	205.48
BLL≥5 within 100m	0.25	184.22
Percent of census tract BLL≥5	2.70	142.93
Social Vulnerability Index in block group, Housing Competition + Disability Theme	0.03	73.69
Social Vulnerability Index in block group, Socioeconomic Theme	-0.04	67.69
1 previous BLL≥5 at birth address	0.79	67.37

Notes: Importance is estimated by the Gini coefficient for the random forest trained to predict BLL≥5µg/dL. All variables are available both for children who were tested and children who were never tested. LASSO coefficient refers to the coefficient in the regression used to reduce the number of parameters, which was cross validated at the minimum mean squared error. BLL denotes Blood Lead level; LASSO: least absolute shrinkage and selection operator.

Table A.2: Logistic Regression Model for Likelihood of BLL≥5µg/dL, Selected Variables

Variable	Coeff	SE	P-val	Reference Category
Mother Age	-0.016	0.003	0.000	
Mother's Education: Grade 8 or Less	0.181	0.106	0.088	Education Unknown
Mother's Education: Grade 9+, No Diploma or Less	0.100	0.102	0.323	Education Unknown
Mother's Education: High School/GED	-0.049	0.101	0.629	Education Unknown
Mother's Education: College <4 Years	-0.160	0.102	0.114	Education Unknown
Mother's Education: College 4 Years	-0.283	0.106	0.007	Education Unknown
Mother's Education: More than College 4 Years	-0.291	0.110	0.008	Education Unknown
Mother Married	0.057	0.022	0.012	
Number of Siblings	0.088	0.007	0.000	
House Built 1900-1909	0.012	0.041	0.768	House Built Before 1900
House Built 1910-1919	-0.022	0.040	0.588	House Built Before 1900
House Built 1920-1929	-0.166	0.040	0.000	House Built Before 1900
House Built 1930-1939	-0.185	0.061	0.002	House Built Before 1900
House Built 1940-1949	-0.374	0.064	0.000	House Built Before 1900
House Built 1950-1959	-0.641	0.052	0.000	House Built Before 1900
House Built 1960-1969	-0.670	0.054	0.000	House Built Before 1900
House Built 1970-1979	-0.630	0.061	0.000	House Built Before 1900
House Built 1980-1989	-0.635	0.084	0.000	House Built Before 1900
House Built 1990-1999	-0.780	0.087	0.000	House Built Before 1900
House Built after 1999	-0.806	0.069	0.000	House Built Before 1900
High Risk Zip Code	0.051	0.028	0.072	
Block Group Median Income	0.000	0.000	0.000	
Percent Rentals in Block Group	-0.496	0.145	0.001	
Percent on Medicaid in Block Group	-0.447	0.125	0.000	
BLL 5+ In Previous Year within 100m	0.325	0.053	0.000	
BLL 5+ In Previous Year within 500m	0.308	0.041	0.000	
BLL 5+ In Previous Year within 1000m	0.123	0.041	0.002	
1 Previous Instance of BLL 5+ at Address	1.215	0.048	0.000	
2 Previous Instances of BLL 5+ at Address	-1.049	0.049	0.000	
1 Previous Instance of BLL 10+ at Address	0.819	0.113	0.000	
2 Previous Instance of BLL 10+ at Address	-1.109	0.116	0.000	
Share of Children with BLL 5+ in Tract	8.154	0.168	0.000	
Share of Children with BLL 10+ in Tract	-3.046	0.388	0.000	

Notes: The table reports coefficient, standard error, and p-value for selected groups of variables that are largely significant in a logistic regression to predict the likelihood that a child had a BLL≥5µg/dL in our training set. The fourth column reports the reference category for categorical variables. The model included also birth year FEs, measures of industrial pollution and proximity to major road, as well as race and ethnicity.

Table A.3: Simulated Effect of Universal Screening by Compliance Rate, Intervention Threshold, and zip Code Risk.

Zip Risk Status	Number of children	Screened Children	Intervention Threshold (µg/dL)	Actual above-threshold BLLs	Target screening rate (percentile)	Additional Children Screened	Predicted additional above-threshold BLLs (Mean (95% CI))
Low	415,365	178,010	5	5,433	50th	79,564	1110 (1174, 1046)
					75th	119,386	1838 (1920, 1756)
					90th	159,620	2622 (2719, 2524)
					100th	237,355	4238 (4361, 4115)
			10	896	50th	79,564	105 (124, 85)
					75th	119,386	181 (207, 155)
					90th	159,620	267 (298, 236)
					100th	237,355	440 (479, 401)
High	319,334	200,257	5	12,682	50th	10,198	428 (466, 390)
					75th	30,851	1398 (1467, 1328)
					90th	59,518	2950 (3051, 2850)
					100th	119,077	6375 (6520, 6230)
			10	2,396	50th	10,198	63 (78, 48)
					75th	30,851	208 (235, 181)
					90th	59,518	437 (476, 398)
					100th	119,077	948 (1004, 892)

Notes: The table estimates the number of additional detected cases of above-threshold BLLs at different rates of compliance with universal screening policy, for intervention thresholds of 5µg/dL and 10µg/dL, stratified by zip code risk status amongst children born in Illinois 2010-2014. Target screening rates were chosen to coincide with the 50^{th}, 75^{th}, 90^{th}, and 100^{th} percentile, of current high-risk zip codes in Illinois where all children should be tested, corresponding to screening rates of 61%, 71%,81%, 100%. Mean and 95% confidence interval are the results of 1000 simulations of testing additional children based on the probability of above-threshold BLLs derived from a calibrated random forest predictor.