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

Ali Abbasi, Francis J. DiTraglia, Ludovica Gazze, Bridget Pals

April 18, 2023

Abstract

Lead exposure still threatens children's health despite policies aiming to identify lead exposure sources. Some US states require *de jure* universal screening while others target screening, but little research examines the relative benefits of these approaches. We link lead tests for children born in Illinois between 2010 and 2014 to geocoded birth records and potential exposure sources. We train a random forest regression model that predicts children's blood lead levels (BLLs) to estimate the geographic distribution of undetected lead poisoning. We use these estimates to compare *de jure* universal screening against targeted screening. Because no policy achieves perfect compliance, we analyze different incremental screening expansions. We estimate that 5,819 untested children had a BLL $\geq 5\mu/dL$, in addition to the 18,101 detected cases. 80% of these undetected cases should have been screened under the current policy. Model-based targeted screening can improve upon both the status quo and expanded universal screening.

Keywords— Lead Poisoning, Environmental Health, Screening, Machine Learning **JEL CODES:** I18; D61.

^{*}Contact: Ali Abbasi: Ali.Abbasi@ucsf.edu, Francis J. DiTraglia: francis.ditraglia@ economics.ox.ac.uk, Ludovica Gazze: Ludovica.Gazze@warwick.ac.uk, Bridget Pals: bridget. pals@nyu.edu. 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, Andrew Oswald, Billy Pizer, David Slusky, Johannes Spinnewjin, and conference participants at APHA for helpful comments and suggestions. All errors are our own.

1 Introduction

A recent literature has emphasized the role of place in shaping early childhood opportunities, however "we know little about the relative importance of the different mechanisms that are typically 'bundled' together within a neighborhood", including pollution (Chyn and Katz, 2021). Lead is a neurotoxic heavy metal that has been widely used 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. Early childhood lead exposure is especially harmful; it is associated with lifelong developmental impacts, including decreased IQ and increased impulsivity and delinquency (Aizer and Currie, 2019; Aizer, Currie, Simon, and Vivier, 2018; Feigenbaum and Muller, 2016; Reyes, 2014; Reyes, 2015; Bellinger, Stiles, and Needleman, 1992; Winter and Sampson, 2017; Grönqvist, Nilsson, and Robling, 2020). These burdens are disproportionately borne by communities of color and families of low socioeconomic status, potentially exacerbating existing inequalities (Zartarian, Xue, Tornero-Velez, and Brown, 2017; Sampson and Winter, 2016).

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 and Urban 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, following the deleading of gasoline between 1973 and 1995 (HUD (U.S. Department of Housing and Urban Development), 2011; Dewalt et al., 2015). In recognition of 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: an estimated 500,000 young children are still poisoned by lead each year in the US (Aizer, Currie, Simon, and Vivier, 2018).

This paper sheds light on one possible reason for this policy failure: imperfect information on the location of lead hazards. A recent study using data from the National Health and Nutrition Examination Survey highlights the potential importance of this channel, finding that US states detect and report only 64% of the actual cases of BLL $\geq 10\mu g/dL$ to the CDC (Roberts et al., 2017). To analyze the role of hidden hazards and undetected lead poisoning, we use machine learning to predict the BLLs of children who were never tested and identify those most at risk of lead exposure. Our results highlight that the spatial distribution of lead exposure sources can be used to better target resources to uncover and remediate hidden lead hazards that would otherwise 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 removal of exposure sources. Deciding which children to screen is thus crucial for identifying lead hazards. Federal guidelines mandate that all children on Medicaid be screened at ages one and two; guidelines for other children vary by state. Importantly, the screening requirement for children on Medicaid is unenforced. Some states, including Illinois, incentivize providers to achieve high screening rates among Medicaid patients as part of bonus schemes that target several performance measures (Tong, Artiga, and Rudowitz, 2022). Fourteen states and the District of Columbia currently adopt *de jure* universal screening, that is testing all children, although their screening rates fall well short of 100% (Michel, Erinoff, and Tsou, 2020). 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 through either socioeconomic and location information or a self-assessment questionnaire (CDC, 1997). Thus, testing is not carried out in response to symptoms, but rather as a preventative measure. This is a reasonable approach given that lead exposure is asymptomatic and occurs very early in life for most of the cases in our sample (CDC, 2013; CDC, 2022).

Currently, the state of Illinois adopts a targeted screening approach. 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 at ages one and two if they reside in one of these high-risk zip codes, if they are on Medicaid, or if they are flagged based 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 were given the option to lower the threshold based on their funding. In 2019, the intervention threshold was lowered to 5µg/dL statewide.

While targeted screening might better balance the costs of screening with the potential benefits of early detection and treatment, its efficacy hinges on the quality of the targeting tools available. At present, these include self-assessment questionnaires (Dyal, 2012; Binns, LeBailly, Fingar, and Saunders, 1999) and existing estimates of the distribution of exposure risks. Precisely because screening is targeted, however, the sample of children used to construct these tools is not representative of the population as a whole and could lead to biased results (Manheimer and Silbergeld, 1998). Moreover, the CDC targeting guidelines were last updated when the intervention threshold was 10µg/dL (Tsoi, C.-L. Cheung, T. T. Cheung, and B. M. Y. Cheung, 2016), so at today's 5µg/dL reference level, or a proposed threshold of 3.5µg/dL, the benefits of targeted screening may diminish (Abbasi, Pals, and Gazze, 2020).

To address these concerns, we propose an improved methodology for predicting undetected childhood lead exposure in Illinois and use it to evaluate a number of alternative targeted screening policies. Our approach combines flexible machine learning tools with a newly-constructed dataset linking lead tests to geocoded birth records and a host of other spatial characteristics thought to predict lead exposure (e.g. housing age, proximity to major roads, and industrial lead emissions). Under a selection-on-observables assumption, these two ingredients allow us to "fill in" the missing BLLs for all children born in Illinois between 2010 and 2014, thus estimating the number of abovethreshold BLLs missed under the current targeted screening policy and the associated costs related to IQ losses these children bear. We use these predictions to estimate how many of these additional above-threshold BLLs would be detected under different policies, taking into account compliance with screening guidelines, as well as different prioritization rules of targeted vs. *de jure* universal screening policies. While ours is not the first paper to use machine learning to predict lead exposure risk, (Lobo, Kalyan, and Gadgil, 2021, Potash et al., 2020), we innovate in several respects. First, our model recognizes that the harm from lead exposure depends on the BLL *itself*, rather than merely whether the BLL exceeds a particular threshold. For this reason, our preferred specification is a continuous-outcome random forest regression model for BLLs. Second, rather than tuning and evaluating our models using default metrics, e.g. root mean squared error, we introduce a novel, policy-relevant metric and use it throughout. Third, we leverage these machine learning results to shed light on the important and understudied question of how best to design a targeted screening policy.

We report two main findings. First, we find evidence of significant under-detection: we estimate that over a quarter of cases of BLLs at or above $5\mu g/dL$ went undetected during our sample period. Second, undetected lead exposure cases appear to be disproportionately located in high-risk zip codes, where all children are supposed to be screened already under the current policy. As a result, increasing screening rates in areas already targeted for screening would uncover more cases than extending *de jure* universal screening at current compliance rates. Improving screening rates in these high-risk zip codes could decrease inequality in human capital and labor market outcomes in Illinois.

This paper contributes to a literature examining the benefits of expanding from targeted to de jure universal screening. So far, studies have projected the benefits of universal screening by multiplying the rate of elevated BLLs among the tested by the number of untested children (Maryland Department of Health and Mental Hygiene, 2015; McMenamin et al., 2018). This approach makes two crucial assumptions. First, it assumes perfect compliance with a hypothetical universal screening program. Based on other evidence from other public health screening programs (Einav et al., forthcoming; Kim and Lee, 2017) this seems unlikely. Second, this approach assumes that elevated BLLs are just as common among untested children as among tested children, in other words that BLLs are "missing completely at random." This assumption seems unlikely to hold a priori, especially given that existing policies target children thought to be at the highest risk. Indeed, our analysis suggests that this assumption is incorrect and that using it dramatically overestimates the benefits of *de jure* universal screening. Proceeding under the weaker assumption of selection-on-observables into testing, also known as "missing at random", we estimated that the rate of BLLs $\geq 5\mu g/dL$ is around *one third* as high among the untested as among the tested. Importantly, our analysis adjusts for a large number of observed characteristics, including those that are currently used to target children for screening. This approach explicitly acknowledges that children tested under a targeted screening scheme are, by construction, a higher risk group.

2 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.¹ 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. We limit birth records to this time period because we use the highest BLL recorded for each child by age two as our outcome of interest. 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. 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. Because venous tests are more reliable than capillary tests, 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 $\leq 3\mu g/dL$ as $3\mu g/dL$). 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 recode these to the mean BLL of children in the same zip and age cohort. We estimate that 6,943 children in our sample had their maximum blood lead level by age two analysed in laboratories with a reporting limit $\geq 5\mu g/dL$.

To obtain information on potential lead exposure sources, 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 (Winkler, 1990). We use addresses at time of birth because these are observed for both screened and unscreened children. As such, we might overestimate the number of unscreened children if they have moved out of state after birth, and we might misallocate children to low- or high-risk zip codes. We successfully geocode birth addresses for 734,699 children, that is 91% of all birth records in our universe are included in our analysis. For each census block group, the American Communities Survey provides data on socioeconomic status, percent homeowners, and social vulnerability index. We obtained data on housing age from the Zillow Transaction and Assessment Dataset, and geocode these data for linkage to birth addresses. We also collected the Environmental Protection Agency's Toxic

¹See Table A1 for more details.

Release Inventory (TRI) data which detail industrial lead emissions by facility, and the location of state and interstate highways from the Illinois Department of Transportation. We then calculate the distance from lead-emitting facilities and roadways to each child's address.

Table 2 shows summary statistics for selected characteristics of children in our sample, stratified by whether a child was tested for lead exposure by age two ad whether they were born in a low or high risk zip code. In both low and high risk zip codes, tested children are more likely to be Black, Hispanic, and have mothers without college education. At birth, they are more likely to live in census block groups with lower median household incomes and in pre-1930 housing.

3 Methodology

We use machine learning methods to predict the incidence of elevated blood lead levels among unscreened children in Illinois under selection-on-observables. In particular, our key maintained assumption is that a child's BLL Y_i is independent of her screening status S_i conditional on the observed covariates X_i given in Table A1. In other words, we assume that

$$Y_i \perp S_i | X_i. \tag{1}$$

As is well-known, (1) cannot be directly tested unless one has access to exogenous instrumental variables or is willing to make parametric functional form assumptions. Given the rich covariate information at our disposal and the flexible models that we employ, however, we consider (1) to be a reasonable approximation. Under (1), we can use the observed BLLs for screened children $(S_i = 1)$ with covariates $X_i = x$ to impute the unobserved BLLs for un-screened children $(S_i = 0)$ with the same covariate values. In this section we explain our methodology for doing precisely this. Section 4 uses the results to carry out policy experiments comparing the effects of alternative screening policies.

There are many possible ways to approach this prediction problem. One potential approach would be to use a regression model to predict BLLs *themselves* based on observed covariates. Another would be us to use a classification model to predict which children are likely to have *elevated* BLLs under the current policy definition, i.e. BLL> 5 μ g/dl. A regression approach makes full use of the information in observed BLLs rather than discretizing based on a single cutoff. On the other hand, BLLs are subject to measurement error and we are primarily interested in detecting children with elevated BLLs. Therefore, a measured BLL above the threshold could be a more reliable indicator than the raw BLL itself; if so, this might favor a classification approach. Because it is unclear *a priori* whether a regression or classification approach will work best in practice, in Section 3.2 we use a holdout sample to choose between a range of regression and classification models.

Precisely because it is unclear whether a classification or regression approach will work best, our methodology is designed to evaluate both classes of model on an equal footing and to accommodate the outputs of either, should it give the best performance in the holdout dataset. To this end, in Section 3.1 we introduce a novel policy-relevant evaluation metric that scores models based on the way that they *rank* children in terms of their risk of elevated BLLs. This metric is compatible with both regression models and with any classification model that outputs a probability. We call the ranking produced under a given model the *risk score* function $r(X_i)$ for that model. This function tells us how best to prioritize children for screening based on their observed characteristics, under the model.

We use the same evaluation metric to score alternative screening approaches in our policy experiments in terms of how well they identify currently unscreened children with high BLLs. While it only requires ordinal information from the *models* under comparison, our evaluation metric accounts for the fact that higher levels of lead exposure are more harmful than lower levels. To do so, it scores the rankings according to the *cardinal* information in observed BLLs from the holdout dataset. In our policy experiments, however, the BLLs of unscreened children are by definition unknown, so to score alternative targeting policies we need to impute the missing BLLs. If the winning model in terms of risk score should happen to be a classification model, we would still need to relate $r(X_i)$ to BLLs to carry out this imputation. Thus, for consistency, we employ a second step that estimates $m(X_i) \equiv \mathbb{E}[Y_i|r(X_i)]$ via local quadratic regression, using data for screened children under our selection-on-observables assumption. While not strictly necessary for a regression model, this second step can nevertheless yield improvements, as we discuss further below.

3.1 Evaluation Metric: Cost-Weighted Targeting Efficiency

To choose an appropriate function $r(\cdot)$ we propose a novel, problem-specific evaluation metric that we call **cost-weighted targeting efficiency** (CWTE). To motivate our approach, it is helpful to begin by considering an alternative that compares screening policies based on false positive and false negative rates, as illustrated by the confusion matrix in Table 1. Suppose that a BLL of 5µg/dl or above is considered "elevated." Then we could view a child with a BLL below 5 who is nonetheless screened as a "false positive" (type II error) and a child with a BLL above 5 who is not screened as a "false negative" (type I error). While natural, this approach has three limitations. First, the false negative rate can always be driven to zero by simply *testing more children*, regardless of how effectively a given policy targets those most at risk. Second, even when comparing two policies that test the same number of children, so that the first criticism does not apply, it is still necessary to choose how much weight to give to each type of error. Policy A could have a lower false positive rate than policy B but a *higher* false negative rate. Third, and most importantly, any hard threshold for BLLs to count as "elevated" is arbitrary: no level of lead in the blood is considered safe and the Environmental Protection Agency has set a maximum contaminant level goal of 0 for lead (EPA, 2021). If a child with a BLL of 4µg/dl is classified as "elevated" is it really reasonable to call this a false positive?

We designed CWTE to address each of these concerns. The intuition is simple: not all elevated blood lead levels are created equal. A BLL of 80 µg/dl is far more harmful, hence more costly, than one of 5µg/dl. For this reason, we assume that policymakers would prefer to identify children with higher BLLs before children with lower BLLs, all else equal. CWTE evaluates a risk score function $r(\cdot)$ by comparing the rankings that it produces to an infeasible optimal ranking that perfectly orders children from highest to lowest BLL. To operationalize this idea, we assign a policy-relevant dollar value to the deviation between a feasible risk ranking and the optimum. To this end, let $c(\cdot)$ be an increasing function that gives the cost associated with a BLL of Y_i . We take this as the value of identifying child *i* as lead-exposed, because of existing evidence suggesting that at least some of these costs can reasonably be averted with appropriate treatment (Billings and Schnepel, 2018).

In our policy experiments below, we construct $c(\cdot)$ using off-the-shelf estimates of the costs of lead exposure at a given level. The per-child social cost of lead exposure would sum health and human capital costs for directly exposed children, including cognitive and non-cognitive losses (Schwartz, 1994), as well as spillovers in terms of lost productivity of parents and disruption in peers' learning (Gazze, Persico, and Spirovska, forthcoming). However, because health costs and indirect costs are harder to estimate for specific BLLs, we follow the literature and focus on cognitive costs (Hollingsworth and Rudik, 2021). Specifically, we take the average IQ point loss per 1µg/dl for different levels of exposure from Lanphear et al., 2005 and Gould, 2009. This is 0.513 for BLLs $\leq 10 \mu g/dl$, 0.19 for BLLs $10 - 19\mu g/dl$, and 0.11 for BLLs $\geq 20 \mu g/dl$. We monetize these losses considering that one IQ point decrease for a three year old is associated with a present value earnings loss of \$20,568 in 2019 dollars (Klemick, Mason, and Sullivan, 2020). We note that a BLL of 1µg/dL — the smallest value that appears in our data — implies an IQ cost of around \$10,500. Figure A1 illustrates the costs associated with each blood lead level. We note, however, that the idea behind CWTE is general: the same approach could be used with *any function* that assigns a cost to BLLs.

Having chosen an appropriate cost function $c(\cdot)$, suppose that we decide to screen n out of a total of N children. The optimal screening policy tests the n children with the highest BLLs Y_i , yielding the highest possible averted cost $C_{\max}(n)$. Screening n children chosen completely at random, on the other hand, yields a total averted cost of $n\mathbb{E}[c(Y_i)]$, on average. Any reasonable risk score should perform better than random screening, but no risk score can perform better than the optimal policy. Let $C_r(n)$ be the total averted cost of risk score function $r(\cdot)$, defined as the sum of $c(Y_i)$ over the BLLs Y_i of the n children with the highest values of $r(X_i)$. Then we have $n\mathbb{E}[c(Y_i)] \leq C_r(n) \leq C_{\max}(n)$.

A natural way to rank two risk score functions, $r_1(\cdot)$ and $r_2(\cdot)$, is to compare their corresponding

averted cost functions $C_1(n)$ and $C_2(n)$ to see which comes closest to the infeasible optimum $C_{\max}(n)$. In practice, however, it is unlikely that policymakers have a specific value of n in mind. Ideally we would prefer a screening policy that performs well over a range of values of n. This complicates the problem because $C_1(n)$ and $C_2(n)$ could cross when plotted as a function of n. Suppose that r_1 is extremely reliable in discerning which children will have a BLL above 20µg/dl but no better than chance at determining which children have a BLL below 5 versus one between 5 and 20. In contrast, r_2 is extremely good at distinguishing BLLs in the range from 1 to 10, but unreliable for larger BLLs. Then we will have $C_1(n) > C_2(n)$ for sufficiently small values of n but $C_1(n) < C_2(n)$ for sufficiently large values of n. This is similar to the problem of comparing machine learning classifiers using operating characteristic (ROC) curves: if the curves cross, the ranking of classifiers depends on the desired false positive rate.

To solve this problem, CWTE *integrates* the averted cost curve, as shown in Figure 2. This idea is analogous to the area under the curve (AUC) measure for classification problems, constructed by integrating the ROC curve. To simplify the figure and computations, we normalize all averted cost curves by $C_{\max}(N)$, the maximum total averted cost if all children were screened, and replace the argument n with n/N, the fraction of children screened. This transformation ensures that averted cost curves always lie within the unit square, and normalizes that of the random screening rule to coincide with the 45-degree line. Finally, we define CWTE for a risk score function $r(\cdot)$ as the area between its averted cost curve $C_r(\cdot)$ and that of random screening, relative to the area between the infeasible optimum averted cost curve and random screening. As such, like the familiar regression R-squared measure, CWTE lies between zero and one, with higher values indicating better performance.

3.2 Choosing a Risk Score Function

To construct an appropriate risk score function $r(\cdot)$ for use in our policy experiments below, we consider a number of alternative machine learning models, including both classification and regression models, before choosing the one with the best out-of-sample CWTE. Since (1) implies

$$\mathbb{E}[Y_i|X_i, S_i = 1] = \mathbb{E}[Y_i|X_i, S_i = 0] \quad \text{and} \quad \mathbb{P}(Y_i \in A|X_i, S_i = 1) = \mathbb{P}(Y_i \in A|X_i, S_i = 0)$$

for any set A, we can use BLLs for children who have been screened, $S_i = 1$, to recover a risk score function that can rank children who have not. In the regression approach, we estimate $r(x) = \mathbb{E}[Y_i|X_i = x, S_i = 1]$; in the classification approach we estimate $\mathbb{P}(Y_i \in A|X_i, S_i = 1)$. We consider three versions of the classification approach, corresponding to different choices of A. The first two set r(x) equal to $\mathbb{P}(Y_i \ge 5|X_i = x, S_i = 1)$ and $\mathbb{P}(Y_i \ge 10|X_i = x, S_i = 1)$, respectively. This approach equates risk score with the probability of having an *elevated* BLL, where "elevated" is defined as ≥ 5 and 10µg/dl. Each of these definitions yields a binary classification problem. As discussed above in Section 3.1, however, the use of a hard binary threshold ignores potentially important distinctions—e.g. 5 versus 80µg/dL—while magnifying unimportant ones—e.g. 4 versus 5µg/dL. For this reason, we consider a third *multi-class* classification approach based on Figure A1. In this approach we calculate the conditional probability given $X_i = x$ that Y_i falls into each of the bins [0, 5), [5, 10), [10, 20), and [20, +∞). To convert these four conditional probabilities into a scalar risk score r(x), we average them with weights equal to the population average BLL within each bin. This is effectively a discrete approximation to the regression approach described above.

A regression or classification model that is too flexible will correct for selection-on-observables (low bias) at the cost of making extremely noisy predictions (high variance). In contrast, a rigidly parametric model will make extremely precise predictions (low variance), but may fail to fully correct for selection-on-observables (high bias). To navigate this trade-off, we tune, estimate, and evaluate all of our models using a training-holdout split. Observations that are used for model estimation are not used for model evaluation, and vice-versa. There is no such thing as "the best" predictive model; there is only the best predictive model *for a given purpose*. For this reason, we explicitly tie our machine learning exercises to the policy question at hand, using the CWTE evaluation metric described above in Section 3.1 to tune each of our competing risk score models and choose which to use in our policy experiments below.

The precise details of our machine learning pipeline are as follows. All of the steps described below are carried out using the R package tidymodels, ensuring that all data processing steps are consistent across models and fully replicable. We begin by constructing the variables listed in Table A1 from our raw data and extracting the subset for which $S_i = 1$, tested children. Note that we exclude the final four variables in the table–distance to provider–from our risk score exercise.² Starting from the observed continuous BLL variable, we construct three categorical BLL variables as follows: an indicator that $Y_i \geq 5$, an indicator that $Y_i \geq 10$, and a categorical variable that indicates which of the bins [0, 5), [5, 10), (10, 20], $(20, +\infty)$ a given observation of a child maximum BLL by age two falls into.

We then construct an 80%-20% training-holdout split for model evaluation, and further subdivide the 80% training sample into five equally-sized cross-validation folds for model tuning. Both the initial training-holdout split and the subsequent cross-validation folds are constructed by sampling randomly within strata defined by the values of the categorical BLL variable. This ensures that the training and holdout data, and each of the cross-validation folds, have the same proportion of BLLs within each of the "bins" listed above. Stratification is crucial for accurate model tuning and evaluation in this example because high BLLs are rare. Without stratification a given cross-validation fold could end up with zero BLLs above 20µg/dl, purely by chance, leading to unreliable

²These are used below to construct a screening propensity score for our policy experiments.

tuning results.

Within strata, we sample observations independently and uniformly at the level of *individual* children rather than geographic aggregates such as zip codes. This choice is based on our selectionon-observables assumption and the goal of this paper. We do not aim to predict BLLs in a new zip code, one for which BLLs are currently unobserved, or a new year, one for which data are not yet available. Instead, we aim to impute the missing BLLs of children who are *already* included in our dataset. In other words, ours is an *interpolation* exercise rather than an *extrapolation* exercise. As a general rule, the design of a cross-validation exercise should mimic the structure of the real prediction problem as closely as possible. If the goal is to predict BLLs for individuals in a new zip code, then cross-validation folds should be constructed by sampling whole zip codes to ensure that the predictive model cannot pick up information from zip code level unobservables. This information would be unavailable when extrapolating to a new zip code, so it should not be used in the tuning and training exercise. In contrast, when the goal is to impute the missing BLLs for children who live in a zip code where we do observe some BLL data, picking up zip code level unobservables is a feature rather than a bug. Our cross-validation exercise simulates this prediction exercise exactly: we randomly drop the BLLs for some children in each zip code, use the others to fill in the gap under selection-on-observables, and check the accuracy of our interpolation against the real data.

After constructing our training-holdout split and cross-validation folds, we define a tidymodels "recipe" for processing the predictor variables. This automatically ensures that all data preparation steps are consistent across sub-samples and models throughout training, tuning, and evaluation. For example, if a predictor is centered around the sample mean, during tuning this mean should only be computed using the appropriate portion of the training data, excluding one of the crossvalidation folds. We encode all categorical predictors from Table A1 as an exhaustive set of dummy variables, including an indicator for "missing" and replace missing values of continuous predictors with the sample mean of the appropriate subset of the training data.

We use random forests to fit the regression and classification models described above, via the R package ranger in concert with tidymodels. Random forests are an attractive choice for this problem because they can approximate complicated non-linearities and interactions between predictor variables in a computationally efficient way, without requiring the user to explicitly construct features that capture these nonlinear effects. They also tend to be quite robust to over-fitting and relatively easy to tune. For the regression model, we use variance as our regression tree splitting rule; for the classification models, binary and multi-class, we use the Gini index.³ For both regression and classification models, we use 500 trees and tune the parameters mtry-the number of variables to consider in each recursive split-and min_n-the minimum number of observations per

³These are the ranger defaults.

leaf-via cross-validation with CWTE as our evaluation metric.⁴ We set all other parameters of ranger to their default values. For each of these random forest models, regression and classification, we construct a parameter grid of 20 combinations of mtry and min_n following a space-filling, Latin hypercube design via tune_grid() and grid_latin_hypercube() from tidymodels. We use the default parameter choices to construct these tuning grids, but double the default number of grid points from 10 to 20 to permit some greater granularity. Overall, our random forest models are quite insensitive to the choice of tuning parameters. Table A2 in the Appendix presents tuning results for our winning random forest regression model, described below.

As an alternative to the random forest models described above, we consider a number of simpler models that do not include interactions or non-linear effects for the predictor variables listed in Table A1.⁵ We include these as a "reality check" to determine if there is anything to be gained from the more complicated, but more flexible, random forest alternatives. As an alternative approach for the continuous random forest model, we consider both plain-vanilla OLS and an elastic-net penalized version of the same. We tune the penalized version via cross-validation, using the CWTE as our evaluation metric. As an alternative approach for the two binary classification models, $\mathbb{E}[\mathbf{1}(Y_i \geq 5)|X_i, S_i = 1]$ and $\mathbb{E}[\mathbf{1}(Y_i \geq 10)|X_i, S_i = 1]$, we additionally consider LASSO-penalized logistic regression, using the glmnet R package. The LASSO-logistic models have a single tuning parameter. Again, we tune this via cross-validation using the CWTE as our evaluation metric. Because glmnet returns parameter estimates along the entire regularization path, i.e. for all values of the LASSO penalty parameter, there is no tuning grid as such. We carry out an exhaustive search over all possible values of the tuning parameter.

Elevated blood levels are a comparatively rare event: fewer than 5% of children in our sample, for example, have a BLL greater than or equal to 5µg/dl. For this reason we estimate two versions of each of our binary classification models, both random forest and LASSO-logistic: one "plain-vanilla" and one in which the dominant class is *downsampled* using step_downsample() from the themis package in tidymodels. We use the default settings for step_downsample(), so that elevated and non-elevated BLLs are equally common in the downsampled data. The cost of balancing the data in this way, of course, is that we lose information for children with non-elevated BLLs. Whether this is a worthwhile tradeoff depends on how the problem of class imbalance interacts with the class of models used for estimation, and the metric used for evaluation. When we incorporate downsampling into our pipeline, all other steps are left unchanged: tuning is still carried out as before, for example. All of this is handled dynamically and automatically via our tidymodels

⁴Note that, while we use CWTE for tuning and model evaluation across all of our specifications, we do not use it as our tree-splitting criterion *within* the classification or regression trees that make up our random forests. While this could be an interesting extension to consider in future work, it would require writing our own custom implementations of the underlying random forest algorithm. Here we prefer to rely on robust, well-tested, off-the-shelf packages for computationally intensive tasks.

⁵We exclude the four distance to provider variables, listed at the bottom of Table A1.

pipeline.

After tuning all of our alternative models using the 80% training sample, we compare their performance on the 20% test sample. Results appear in Table A3. To provide a benchmark for our machine learning results, Table A4 presents the out-of-sample CWTE for a number of "naïve" models that do not use covariates or machine learning. For example, prioritizing children in the holdout sample by the average IQ cost of children from the same zip code in the training sample, yields a CWTE of 0.28. From a comparison of these tables, three features stand out. First, all of the machine learning models from Table A3 clearly outperform the "naïve" models from Table A4. Second, downsampling gives mixed results. While it improves the performance of models that predict the probability that a child has a $BLL \geq 10 \mu g/dL$, it worsens the performance of those that predict the probability that a child has a BLL $\geq 5\mu g/dL$. Third, the continuous outcome random forest regression model, in which $r(x) = \mathbb{E}[Y_i|X_i = x, S_i = 1]$ is the clear winner in terms of CWTE, with a value of around 0.39, outperforming the runner up elastic net model. Figure A2 in the Appendix plots the proportion of children in the holdout set with $BLL \geq 5\mu g/dL$ in each vigintile (5% bin) of risk score as predicted by our winning model. Notably, the lowest 70% of children in the holdout sample as ranked by our risk score have are less likely to have BLLs 5+ than the average child in the holdout sample. Very few children in the bottom quartile have elevated BLLs. Because this is an out-of-sample exercise, it suggests that our model discriminates well between low- and high-risk children. For the remainder of the paper we will take r(x) to be the estimate of $\mathbb{E}[Y_i|X_i = x, S_i = 1]$ obtained by *re-estimating* the winning regression random forest model using the data for all tested children with tuning parameters set according to the cross-validation exercise described above.

To give a sense of the covariates from Table A1 that are particularly relevant for predicting Y_i , Table A5 presents variable importances for our top-performing model. These are computed using **ranger** with the **importance =** '**impurity**' option. Every time a particular predictor variable is used to make a recursive split in one of the regression trees that make up our random forest, this improves the in-sample predictive MSE. The impurity variable importance measure *averages* these improvements from a given variable across all trees, and then makes relative comparisons across variables. On the whole, the most important variables are at the neighborhood-level: measures of the prevalence of lead exposure, characteristics of the housing stock, and socio-economic variables. We caution that this pattern should not necessarily be given a causal interpretation. Moreover, it does not imply that individual-level variables have no predictive power. The age of any given house is likely highly correlated with the age of the houses that surround it, for example. This kind of multicollinearity presents no problem for machine learning prediction, but it does mean that variable importances should be taken with a grain of salt.

Notably absent from Table A5 are measures of lead exposure sources outside the home, including distance to major roads, and industrial lead emissions reported to the Toxic Release Inventory

(TRI). Bearing in mind the point about multicollinearity mentioned above, we conjecture that the low predictive power of these variables may be explained by the following factors. First, that proximity to major roads is not as strong of a predictor of lead exposure is perhaps less surprising in light of the recency of our sample (birth cohorts 2010-2014). Indeed, Aizer and Currie, 2019 find that the relationship between a child's BLL and traffic on roads within 50 meters of the child's home for children born in the early 1990s but not for those born in 2004. The relationship between road proximity and child BLLs has likely attenuated over time because the amount of lead in soil has declined following the deleading of gasoline between 1979 and 1986. As for industrial lead emissions, we are unaware of a literature discussing attenuation over time. Still, we see two plausible explanations for the low predictive power of TRI lead emissions. First, Hollingsworth and Rudik, 2021 note that lead from NASCAR races appears to travel up to 50 miles away. We look at plants within two kilometers of children's homes, and most lead is released through stacks. Thus, it could be that there are no differences at these close distances. Second, lead emissions from TRI facilities might be generating clusters of lead-exposed children, something that would be picked up by the neighborhood-level counts of lead-exposed children variables we include in our model, an example of the multicollinearity concern.

3.3 Policy Experiments

Our main policy experiments are counterfactuals in which some of the unscreened children from our dataset are shifted into screening while all currently-screened children remain so. Each experiment corresponds to a different way of deciding which currently-unscreened children are shifted. We consider a number of possibilities including targeting based on zip code, targeting based on the risk score measure $r(\cdot)$ from Section 3.2, and targeting based on the screening propensity score $p(x) \equiv \mathbb{P}(S_i = 1|X_i = x)$. Because we use a regression-based approach rather than a propensity-score weighting approach to address selection-on-observables, we require an estimate of the propensity to be screened, p(x), for this latter policy experiment. We construct this propensity score via LASSO-penalized logistic regression, where X_i contains all of the variables from Table A1, including the four distance-to-provider variables.⁶

We score alternative screening policies by the total averted cost generated by shifting the relevant children under each policy, using the IQ cost function from Figure A1 and predicted BLLs based on covariates. We also construct intervals that show how these averted cost figures vary over the range of predictive uncertainty in BLLs given observed covariates. Because we observe all BLLs

⁶Our machine learning pipeline for the propensity score model is identical to that of the other LASSOlogistic models from Section 3.2 with two exceptions. First, because there are approximately equal shares of screened and unscreened children, there is no need for stratification when constructing training-holdout and cross-validation splits and no need for downsampling in estimation. Second, because the outcome variable for this regression is S_i , whether or not a child is screened, CWTE is inapplicable, so we use AUC as our evaluation metric.

for the population of tested children, by definition, these uncertainty intervals treat r(x) as fixed. Uncertainty arises because r(x) does not perfectly predict the BLL of a child with $X_i = x$. We quantify this predictive uncertainty using a simple bootstrap-based procedure described below.

Our winning risk score function r(x) from 3.2 is the random forest approximation to the conditional mean function $\mathbb{E}[Y_i|X_i = x, S_i = 1]$. This approximation does the best job of prioritizing children for screening in the out-of-sample holdout data, using CWTE as the evaluation metric. While we could in principle use r(x) directly as our BLL prediction for unscreened children with $X_i = x$, for consistency with models in which $r(X_i)$ is not a regression function, we employ a second step local quadratic regression of Y_i on $r(X_i)$, using data for the tested. Figure 3 plots the resulting regression curve (solid) along with the 45-degree line (dashed). As seen from the figure, the winning random forest regression model slightly over-predicts very low BLLs and under-predicts very high BLLs. We speculate that under-prediction in the right tail results from the rarity of extremely high BLLs in our sample and the fact the random forest model uses RMSE as the tree-splitting criterion. Together these two factors may imply that the model is insufficiently "rewarded" for distinguishing between very high BLLs. The slight over-prediction of low BLLs could potentially relate to detection thresholds and measurement error, or merely be a mechanical compensation for the over-prediction of high BLLs. Fortunately both the over- and under-prediction problems are completely resolved by our second step, which effectively "re-calibrates" the risk score r(x). Because $m(\cdot)$ is monotone, this has no implications for the order in which children should be screened. given their covariates.

In our policy exercises, for an untested child with covariates $X_i = x$ we predict that $Y_i = m(x)$ where $m(X_i) \equiv \mathbb{E}[Y_i|r(X_i), S_i = 1]$ is the regression from 3. To incorporate predictive uncertainty into our policy analysis, we use a simple bootstrap-based procedure that is justified under our selection-on-observables assumption. We approximate the distribution of $Y_i|X_i = x$ among the unscreened by computing the residuals from Figure 3 for the 50 tested children whose risk scores are *closest* to r(x). We then sample from these residuals and add them to m(x) to approximate the distribution of $Y_i|X_i = x$. This allows us to capture the heteroskedasticity and asymmetry evident from Figure 3. In our policy experiments, we use 1000 bootstrap samples to construct uncertainty intervals for the value of each policy.

4 Results

In our sample, 18,101 tested children have a BLL $\geq 5\mu g/dL$ and 3,290 tested children have a BLL $\geq 10\mu g/dL$. We estimate substantial underdetection of lead exposure: among children born between 2010-2014, current testing practices detected 76% of cases of BLL $\geq 5\mu g/dL$ and 82% of cases of BLL $\geq 10\mu g/dL$. Indeed, our model predicts an additional 5,819 (95%CI 5,696-5,940) of the

349,489 untested children had BLL $\geq 5\mu g/dL$ (Table 3). We also predict an additional 699 (95%CI 660-741) of the untested children had BLL $\geq 10\mu g/dL$.

To investigate where the hidden costs of children with undetected BLLs are highest, Figure 4 plots the distribution of costs from IQ losses related to lead exposure of untested children in highand low-risk zip codes. The distribution of "missed" costs of lead exposure for high-risk zip codes is shifted to the right with respect to the one for low-risk zip codes, suggesting the most severe undetected poisoning cases appear to be concentrated in areas already identified as high risk. This finding is striking as these children should have been tested under Illinois' existing screening policy. Table 3 further illustrates this point by showing the number of and costs associated with undetected cases of above-thresholds BLL in high-risk zip codes. While there are fewer untested children in high-risk zip codes (112,997 vs. 236,492), 80% of untested children with predicted BLL \geq 5µg/dL lived in high-risk zip codes (4,671) rather than low-risk ones (1,148). Our model also predicts that 82% of children with undetected BLL \geq 10µg/dL were in high-risk zip codes (576) rather than low-risk zip codes (123).

This unequal distribution of lead hazards, which appear concentrated in high-risk zip codes, suggests that targeted screening has merit. Figure 5 investigates whether our model could be used to improve targeting via an out-of-sample exercise using the 20% holdout sample. Each panel plots the distribution of costs from IQ losses across zip codes, comparing low-risk (pink) against high-risk (blue). The left panel uses the official IDPH definition of risk, while the right panel uses an alternative definition based on our preferred random forest model, fitted to the 80% training sample. In particular, we designate the 507 zip codes with the highest average risk scores as "highrisk", matching the number of IDPH-designated high-risk zip codes in the holdout sample.⁷ Based on the holdout data, we see that children in IDPH-designated high-risk zip codes experience greater harm from lead exposure on average. For example, the median high-risk zip code has an average cost per child of \$18,600 compared to \$16,100 for low-risk zip codes (left panel). As seen in the right panel of the figure, however, the risk scores estimated by our model from Section 3.2 perform better, yielding a designation of 507 high-risk zip codes with a median cost of \$20,300 per child compared to \$15,800 for our model-designated low risk zip codes. Thus, by adjusting the definition of high-risk zip codes, which would involve reassigning 35% of high-risk zip codes in Illinois (Table 4), states may be able to detect a higher number of above-threshold BLLs without increasing the number of high-risk zip codes.

The preceding exercise suggests that there may be gains from re-assigning some zip codes from high- to low-risk and vice-versa. A similar exercise could be carried out at the individual level, by asking whether our model can accurately identify low risk children among the currentlyscreened solely based on the values of their covariates X_i . If so, policymakers could potentially

⁷Children in the holdout sample cover 507 out of the 580 official high-risk zipcodes and 1,252 of the 1,364 Illinois zip codes in our sample.

determine which children should *not* be prioritized for screening. Figure 6 illustrates the out-ofsample performance of our risk score by plotting average monetized IQ losses in the holdout sample against risk scores computed from our preferred random forest model fitted to the training data. Each point in the plot is a 5% bin of risk score moving from highest to lowest risk along the xaxis. We see that the average cost declines very rapidly as risk scores fall. This suggests that, if policymakers desired to identify groups of currently-tested children who are at particularly low risk of serious adverse consequences from lead exposure, they could do so reliably out-of-sample based on our model. Note that this exercise, unlike the others that we discuss in this section, does *not* rely on selection-on-observables, because it is solely based on data from children who were in fact screened. More broadly, this figure demonstrates the value of targeting based on a model such as ours, compared to universal screening.

Next, we investigate the role of compliance with screening guidelines in averting lead exposure costs. To do so, Table 5 considers two scenarios.⁸ In Panel A, we examine the effect of increasing testing rates in zip codes with low testing rates to the level of screening compliance of the median, 75th, and 90th percentile high-risk zip code under the current targeting system. These correspond to screening rates of 61.1%, 71.4%, and 81.3% (Figure A3). In Panel B, we consider raising the overall screening rate in Illinois to those same levels, irrespective of zip code of residence. Scenario A takes it as given that targeting will be carried out at the zip code level–the current practice–while scenario B imagines targeting in a centralized way across the state. We would expect Scenario B to yield better individualized targeting, but perhaps at the cost of additional logistical complications.

The question then arises of which additional children would be tested under each scenario. We evaluate the screening policies under different priority systems: randomly sampling among untested children, screening based on the propensity scores—highest or lowest—from Section 3.3, considering children with high vs. low lead exposure risk scores as predicted by our preferred regression random forest model, and considering children who are closer vs. farther away from screening providers, a factor that has been shown elsewhere to influence the likelihood of getting tested (Gazze, 2022). Importantly, we remain agnostic as to the logistical feasibility of each prioritisation system. In other words, we do not attempt to quantify the costs of prioritizing children with the highest risk scores to be screened. Moreover, we do not assume that prioritizing based on distance to provider is easier. Instead this exercise only considers the correlation between distance to provider and predicted BLLs. Figure 7 shows that the correlation between predicted BLLs and screening propensity scores is non-monotonic, and therefore it is plausible that targeting based only on risk scores might not yield the desired screening rates. For each different policy, we report the average averted cost per newly-tested child—this reflects costs for children with BLL< 5µg/dL—as well as

⁸Table 5 uses predictions from our preferred model. Results are qualitatively similar for the runner-up model, the Lasso-penalized logistic model that predicts the probability that a child has a BLL $\geq 5\mu g/dL$.

the same average restricted to those whom we predict to have $BLLs \ge 5\mu g/dL$.

Panel A of Table 5 shows that a *de jure* universal screening regime such that all zip codes achieved the same screening compliance as the median high-risk zip code under Illinois' existing regime could avert between \$15,209 and \$15,403 per child screened in costs associated with IQ losses from lead exposure, depending on the prioritization rule (confidence intervals are very tight). If we assign an averted cost of zero to children with BLLs below $5\mu g/dL$, the averted cost per child screened ranges from \$1,033 to \$1,126. These values are policy-relevant because, given the current intervention threshold of 5 μ g/dl, children with lower BLLs would not be eligible to receive treatment or other follow-up interventions. Notice that the averted cost falls by over a factor of ten when we exclude children with BLLs below 5 $\mu g/dl$ from consideration. This is because the overwhelming majority of children with detectable levels of lead in their blood have BLLs below 5 µg/dl. Moreover, our IQ cost function is concave so the largest marginal effect occurs at lower BLLs. Interestingly, because screening rates are lowest in low-risk zip codes (Figure 8), prioritization rules that operate within a given zip code do not appear to make much of a difference relative to random screening. Nevertheless, prioritization based on our risk score measure yields the highest benefits when looking only at children with BLLs above $5\mu g/dL$ and prioritization based on farthest distance to providers yields the highest benefits overall, while prioritization based on closest distance yields the smallest benefits. Raising compliance to the 75th or the 90th percentile of current high risk zip codes would only result in average averted costs between \$15,527 and \$15,727 and between \$15,663 and \$15,769, respectively. On the whole, increasing screening rates in every zip code to match that of the top high-risk zip codes provides benefits only insofar as it leads to detecting more cases of lead exposure in zip codes already designated as high-risk.

Whereas panel A of Table 5 considers the effect of testing additional children in zip codes with screening rates below the median zip code, Panel B considers the effect of raising the overall screening rate in the state of Illinois to the same level by prioritizing children *across* zip codes. In this exercise, targeting provides considerable benefits over randomly sampling additional children to test.⁹ Regardless of whether we score all BLLs according to Figure A1 or treat those below 5µg/dL as zeros, averted costs are highest when children are prioritized based on our risk score measure from Section 3.2 (26,370 and \$7,659 respectively when screening 61% of children, the median screening rate in high-risk zip codes). It is also worth noting that in this case prioritizing children with the highest screening propensity or those that are closest to providers also yields higher benefits than random sampling, suggesting that there is a relationship between predicted BLLs and propensity or ability to be screened that could be leveraged with the right policies (see

⁹Note that, when prioritizing children across the entire state rather than within zip codes, random screening by definition produces the same average averted cost at the 50%, 75% and 90% screening compliance rates. When prioritizing within zip codes, however, this is no longer the case because changing the compliance rate changes the set of zip codes from which the random sample is drawn.

e.g., Figure 7).

Together, these estimates suggest that careful targeting is crucial for the effectiveness of any expansion of existing childhood lead screening. Screening need not be universal to be effective, and *de jure* universal screening that is not *de facto* universal may be ineffective. A potential benefit of *de jure* universal screening that our analysis does not capture is lower communication and logistical costs. For example, providers would not need to check a child's zip code of residence under a universal screening policy, and this might increase compliance. Such gains in compliance, however, are far from assured. All zip codes in Chicago are high-risk, implying that every child in the city should be screened. If this city-wide universal screening policy lowered communication costs and increased compliance, we would expect to see higher compliance in Chicago compared to other high-risk zip codes that are more dispersed. But this does not appear to be the case: average screening rates in Chicago were 65%, compared to 64% in high-risk zip codes outside of Chicago.

In previous work some of us have shown that the relative importance of different lead exposure sources shifts as the intervention threshold is lowered, which may make it more difficult to identify children with elevated BLLs using proxies for lead exposure (Abbasi, Pals, and Gazze, 2020). This pattern may explain why we estimate more than 8 times as many *undetected* cases of BLL \geq 5µg/dL relative to \geq 10µg/dL when there were approximately 5.5 times as many *observed* cases of BLL \geq 5µg/dL as BLL \geq 10µg/dL during the study period. In spite of this, the current definition of high-risk zip codes appears to cover most undetected cases of both BLL \geq 5µg/dL and BLL \geq 10µg/dL cases. Because a higher share of BLL \geq 5µg/dL cases than BLL \geq 10µg/dL cases go undetected, our results highlight the importance of relying on data to identify children most at-risk within areas already flagged for higher screening.

As we have seen from Table 5, the benefits of targeted screening are large. We would like to assess whether increasing screening rates among the currently-untested using our risk score measure passes the cost-benefit test. Carrying out a full cost-benefit analysis, however, requires estimates of the *costs* of alternative screening policies as well. The price of private tests in Illinois ranges up to $43.^{10}$ This gives a rough approximation of the marginal direct cost of testing an additional child. Targeting children based on risk also increases the direct costs of a screening program, by requiring additional data linkages and analysis. These costs are fixed with respect to the number of children tested, but recur on a yearly basis as new data become available. We would estimate that these costs are modest: perhaps in the range of a few thousand dollars of an analyst's time per year. There may also be additional logistical and communication costs, although these are hard to quantify. It is even more challenging to estimate the *indirect* costs of screening, such as the opportunity cost of time for parents–these include travel costs to the doctor's office and health care service providers–and non-monetary costs, e.g., pain if a venous blood sample is taken (Gazze,

¹⁰See https://www.luc.edu/media/lucedu/hhhci/pdf/leadsafeil/LeadSafeILDirectory061_.pdf, accessed in November 2021).

2022). As a back-of-the-envelope calculation, suppose that we only assign a positive benefit to children with BLLs above $5\mu g/dL$ -a very conservative assumption. Then we see from Table 5, that the average benefit per child of expanding screening from the current rate of 49% to 61% using our risk score measure is \$7,659 (Panel B – IQ Cost if BLL 5+ / Risk Score Top). It seems implausible that total screening costs per child could even come close to this number. Even a cost of \$1,000 per child seems very high. For this reason, we consider it plausible that improving targeted screening using our model passes the cost-benefit test. Because targeted screening appears to have higher average benefits than *de jure* and *de facto* universal screening, and because achieving 100% universal screening would be at least as costly as targeted screening, it appears that targeted screening should be more cost-effective than universal screening.

5 Conclusion

We estimate the extent and geographic distribution of undetected lead poisoning in Illinois using administrative data and machine learning tools. We find that current testing practices failed to detect 24% of BLL $\geq 5\mu$ g/dL and 18% of BLL $\geq 10\mu$ g/dL among children born between 2010 and 2014. Moreover, 80% of Illinois children with undetected BLL $\geq 5\mu$ g/dL lived in designated highrisk zip codes where every child should already be tested under Illinois' current testing guidelines. The state defines these zip codes as high-risk based on the age of their housing and the relatively low socioeconomic status of their residents. This suggests 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 terms of averted lead exposure costs from improving compliance with existing zip code-targeted screening policies, rather than expanding to a *de jure* 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 (Gazze, 2022). Still, we find that predicted BLLs correlate positively with both proximity to providers and predicted screening propensity, suggesting that low-cost interventions might shift some of these high-risk children into screening.

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 BLL. Our risk score function could be used to categorize zip codes as high-risk in a targeted screening program. Indeed, by adjusting the definition of high-risk zip codes using our risk scores, which would involve reassigning 35% 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. Moreover,

we find that an individual approach to targeting, one that is based on each child's risk score rather than the model-based high-risk zip code definition, would achieve even higher benefits. Importantly, these risk scores could also be used to educate providers and patients about their risk and encourage proactive home inspections, although response rates have been low (Potash et al., 2020).

Our approach could be adapted for other states to inform lead testing policy, evaluate the effects of changing intervention thresholds, and identify the children at highest risk for lead exposure. Further extensions of the model could add data on additional pathways for lead exposure, such as lead in drinking water or toys, and parental occupational exposure. However, we note that housing vintage likely partially accounts for the effects of lead in water because the use of lead pipes and service lines follows historical patterns (Rabin, 2008). Additionally, the missing exposure sources are understood to represent only a small part of total lead exposure (Zartarian, Xue, Tornero-Velez, and Brown, 2017).

References

- Abbasi, Ali, Bridget Pals, and Ludovica Gazze (2020). "Policy Changes and Child Blood Lead Levels by Age 2 Years for Children Born in Illinois, 2001–2014". In: American Journal of Public Health 0, e1–e7.
- Aizer, Anna and Janet Currie (2019). "Lead and juvenile delinquency: new evidence from linked birth, school, and juvenile detention records". In: *Review of Economics and Statis*tics 101.4, pp. 575–587.
- Aizer, Anna, Janet Currie, Peter Simon, and Patrick Vivier (2018). "Do low levels of blood lead reduce children's future test scores?" In: American Economic Journal: Applied Economics 10.1, pp. 307–41.
- Bellinger, David C, Karen M Stiles, and Herbert L Needleman (1992). "Low-level lead exposure, intelligence and academic achievement: a long-term follow-up study". In: *Pediatrics* 90.6, pp. 855–861.
- Billings, Stephen B. and Kevin T. Schnepel (July 2018). "Life after Lead: Effects of Early Interventions for Children Exposed to Lead". In: American Economic Journal: Applied Economics 10.3, pp. 315–44.
- Binns, Helen J, Susan A LeBailly, Ann R Fingar, and Stephen Saunders (1999). "Evaluation of risk assessment questions used to target blood lead screening in Illinois". In: *Pediatrics* 103.1, pp. 100–106.
- Centers for Disease Control and Prevention (1997). Screening young children for lead poisoning: guidance for state and local public health officials. ERIC Clearinghouse.
- (2013). "Blood Lead Levels in Children Aged 1–5 Years United States, 1999–2010." In: Morbidity and Mortality Weekly Report.
- (2022). "Health Effects of Lead Exposure." In.
- Chyn, Eric and Lawrence F Katz (2021). "Neighborhoods matter: Assessing the evidence for place effects". In: *Journal of Economic Perspectives* 35.4, pp. 197–222.
- Dewalt, F Gary, David C Cox, Robert O'Haver, Brendon Salatino, Duncan Holmes, Peter J Ashley, Eugene A Pinzer, Warren Friedman, David Marker, Susan M Viet, et al. (2015). "Prevalence of lead hazards and soil arsenic in US housing". In: Journal of environmental health 78.5, pp. 22–29.
- Dyal, Brenda (2012). "Are lead risk questionnaires adequate predictors of blood lead levels in children?" In: *Public Health Nursing* 29.1, pp. 3–10.

- Einav, Liran, Amy Finkelstein, Tamar Oostrom, Abigail J Ostriker, and Heidi L Williams (forthcoming). "Screening and Selection: The Case of Mammograms". In: American Economic Review.
- EPA (2021). "National primary drinking water regulations: Lead and copper rule revisions".In: *Fed. Regist.* 84.219.
- Feigenbaum, James J and Christopher Muller (2016). "Lead exposure and violent crime in the early twentieth century". In: *Explorations in economic history* 62, pp. 51–86.
- Gazze, Ludovica (2022). "Hassles and Environmental Health Screenings: Evidence from Lead Tests in Illinois". In: *Journal of Human Resources*, 0221–11478R2.
- Gazze, Ludovica, Claudia Persico, and Sandra Spirovska (forthcoming). "The long-run spillover effects of pollution: How exposure to lead affects everyone in the classroom". In: *Journal* of Labor Economics.
- Gould, Elise (2009). "Childhood lead poisoning: conservative estimates of the social and economic benefits of lead hazard control". In: *Environmental health perspectives* 117.7, pp. 1162–1167.
- Grönqvist, Hans, J. Peter Nilsson, and Per-Olof Robling (2020). "Understanding How Low Levels of Early Lead Exposure Affect Children's Life Trajectories". In: Journal of Political Economy 128.9, pp. 3376–3433. DOI: 10.1086/708725.
- Hollingsworth, Alex and Ivan Rudik (2021). "The effect of leaded gasoline on elderly mortality: Evidence from regulatory exemptions". In: American Economic Journal: Economic Policy 13.3, pp. 345–73.
- HUD (U.S. Department of Housing and Urban Development) (2011). American Healthy Homes Survey Lead and Arsenic Findings.
- Kim, Hyuncheol Bryant and Sun-mi Lee (2017). "When public health intervention is not successful: Cost sharing, crowd-out, and selection in Korea's National Cancer Screening Program". In: Journal of health economics 53, pp. 100–116.
- Klemick, Heather, Henry Mason, and Karen Sullivan (2020). "Superfund cleanups and children's lead exposure". In: Journal of environmental economics and management 100, p. 102289.
- Lanphear, Bruce P., Richard Hornung, Jane Khoury, Kimberly Yolton, Peter Baghurst, David C. Bellinger, Richard L. Canfield, Kim N. Dietrich, Robert Bornschein, Tom Greene, Stephen J. Rothenberg, Herbert L. Needleman, Lourdes Schnaas, Gail Wasserman, Joseph Graziano, and Russell Roberts (Mar. 2005). "Low-Level Environmental Lead Exposure and Children's Intellectual Function: An International Pooled Analy-

sis". In: *Environmental Health Perspectives* 113.7, pp. 894–899. ISSN: 0091-6765. DOI: 10.1289/ehp.7688.

- Lobo, GP, B Kalyan, and AJ Gadgil (2021). "Predicting childhood lead exposure at an aggregated level using machine learning". In: International Journal of Hygiene and Environmental Health 238, p. 113862.
- Manheimer, Eric W and Ellen K Silbergeld (1998). "Critique of CDC's retreat from recommending universal lead screening for children." In: *Public health reports* 113.1, p. 38.
- Maryland Department of Health and Mental Hygiene (2015). Maryland targeting plan for children areas at risk for childhood lead poisoning.
- McMenamin, Sara B, Sarah P Hiller, Erin Shigekawa, Troy Melander, and Riti Shimkhada (2018). "Universal lead screening requirement: a California case study". In: American journal of public health 108.3, pp. 355–357.
- Michel, Jeremy J, Eileen Erinoff, and Amy Y Tsou (2020). "More Guidelines than states: variations in US lead screening and management guidance and impacts on shareable CDS development". In: *BMC Public Health* 20.1, pp. 1–10.
- Potash, Eric, Rayid Ghani, Joe Walsh, Emile Jorgensen, Cortland Lohff, Nik Prachand, and Raed Mansour (2020). "Validation of a machine learning model to predict childhood lead poisoning". In: JAMA network open 3.9, e2012734–e2012734.
- Rabin, Richard (2008). "The lead industry and lead water pipes "A Modest Campaign"".In: American journal of public health 98.9, pp. 1584–1592.
- Reyes, Jessica Wolpaw (2014). "The social costs of lead". In: Lead: The global poison humans, animals, and the environment. May, pp. 1–4.
- (July 2015). "Lead exposure and behavior: Effects on antisocial and risky behavior among children and adolescents". In: *Economic Inquiry* 53.3, pp. 1580–1605. ISSN: 00952583. DOI: 10.1111/ecin.12202.
- Roberts, Eric M, Daniel Madrigal, Jhaqueline Valle, Galatea King, and Linda Kite (2017). "Assessing child lead poisoning case ascertainment in the US, 1999–2010". In: *Pediatrics* 139.5.
- Sampson, Robert J and Alix S Winter (2016). "The racial ecology of lead poisoning: Toxic inequality in Chicago neighborhoods, 1995-2013". In: Du Bois Review: Social Science Research on Race 13.2, pp. 261–283.
- Schwartz, J (1994). "Societal benefits of reducing lead exposure". In: *Environmental Research* 66, pp. 105–124.

- Tong, Michelle, Samantha Artiga, and Robin Rudowitz (2022). *Mitigating Childhood Lead Exposure and Disparities: Medicaid and Other Federal Initiatives*. Tech. rep. Kaiser Family Foundation.
- Tsoi, Man-Fung, Ching-Lung Cheung, Tommy Tsang Cheung, and Bernard Man Yung Cheung (2016). "Continual decrease in blood lead level in Americans: United States National Health Nutrition and examination survey 1999-2014". In: *The American journal* of medicine 129.11, pp. 1213–1218.
- Winkler, William E (1990). "String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage." In.
- Winter, Alix S and Robert J Sampson (2017). "From lead exposure in early childhood to adolescent health: A Chicago birth cohort". In: American journal of public health 107.9, pp. 1496–1501.
- Zartarian, Valerie, Jianping Xue, Rogelio Tornero-Velez, and James Brown (Sept. 2017). "Children's Lead Exposure: A Multimedia Modeling Analysis to Guide Public Health Decision-Making". en. In: *Environmental Health Perspectives* 125.9, pp. 097009 1–10. ISSN: 0091-6765, 1552-9924. (Visited on 12/10/2018).

Figures



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

Notes: The figure plots the zip codes currently classified as high-risk according to guidelines by the Illinois Department of Public Health.



Figure 2: Constructing Cost-Weighted Targeting Efficiency

Notes: The dotted curve gives the averted cost of the infeasible optimal screening rule as a function of the share of children screened; the dashed line gives the value (in expectation) of random screening. The solid curve gives the value C_r of a feasible screening policy based on risk score $r(\cdot)$. CWTE equals the area of the gray shaded region divided by the area between the dotted curve and dashed line. The averted cost from screening all children is normalized to one.



Figure 3: Observed and Estimated BLLs

Notes: The figure plots a local quadratic regression of observed BLLs among the tested on risk scores from the continuous random forest model (solid line), and comparing it to the 45-degree line (dashed line).

Figure 4: Model-implied Lead Exposure Costs (IQ Losses) per Untested Child in High- and Low-Risk Zip Codes



IQ Cost in High and Low Risk Zip Codes (log 10 scale)

Notes: The figures plot the unnormalized distribution of IQ costs of untested children in high-and low-risk zip codes based on the official definition of zip code risk, imputing BLLs using our preferred random forest model fitted to the data for all tested children. The relative areas under each curve correspond to the relative number of zip codes in each group. Vertical lines indicate the median costs for the two groups rounded to the closest 100.

Figure 5: Lead Exposure Costs (IQ Losses) per Child in High- and Low-Risk Zip Codes, Official vs. Model-Based Definition Evaluated in the Holdout Sample



Notes: The figures plot the unnormalized distribution of IQ costs of children in the holdout sample across high-and low-risk zip codes based on the official (left) and model-based definition of zip code risk (right). The relative areas under each curve in a given panel correspond to the relative number of zip codes in each group. The model-based definition designates the top 507 zip codes by average IQ cost as high risk, the same number as the official definition in the holdout sample. Vertical lines indicate the median costs for the two groups rounded to the closest 100.

Figure 6: Average IQ Costs by Risk Score in the Holdout Sample



Average IQ Costs of Highest Risk Children

Notes: The x-axis plots rank in the risk score distribution from highest to lowest risk, based on our preferred random forest model fitted using the training data. Each point represents a 5% bin of the risk score variable. The y-axis plots average IQ costs calculated from the 20% holdout sample.

Figure 7: Correlation between Predicted Screening Probability and Predicted BLL



Predicted BLL by Screening Probability 5th Degree Polynomial Regression

Notes: The figure plots a fifth-degree polynomial regression of predicted BLLs on screening propensity scores for untested children.

Figure 8: Screening Rates in Low- and High-Risk Zip Codes



Share of Children Tested by Zip Code: High and Low Risk Zips IDPH Risk Definition

Notes: The figure plots the distribution of screening rates in low- and high-risk zip codes for Illinois children born between 2010 and 2014.

Tables

	Elevated BLL?		
	No	Yes	
Unscreened Screened	True — False +	False – True +	

 Table 1: Confusion Matrix for Evaluating a Hypothetical Screening Policy

Notes: The table reports the four possible outcomes of a screening policy in terms of whether a child was screened (row dimension) and whether the child has an above-threshold BLL (column dimension).

	Low Risk		High Risk		
	Unscreened Screened		Unscreened	Screened	
	n = 236492	n = 178873	n = 112997	n = 206337	
Black	0.08 (0.28)	0.13 (0.33)	0.24 (0.43)	0.28 (0.45)	
Hispanic	$\begin{array}{c} 0.13 \\ (0.33) \end{array}$	0.23 (0.42)	0.22 (0.42)	$\begin{array}{c} 0.33 \\ (0.47) \end{array}$	
Teen Mother	$0.04 \\ (0.19)$	0.08 (0.27)	0.08 (0.27)	0.11 (0.31)	
Single Mother	$0.24 \\ (0.43)$	$0.43 \\ (0.49)$	$0.43 \\ (0.49)$	$0.54 \\ (0.50)$	
Mother Education: High School or Less	$0.23 \\ (0.42)$	$\begin{array}{c} 0.39 \\ (0.49) \end{array}$	$\begin{array}{c} 0.37 \\ (0.48) \end{array}$	0.48 (0.50)	
Median Income in Block Group	73465.78 (30592.48)	62282.59 (28101.09)	54509.30 (28896.32)	$\begin{array}{c} 48074.19 \\ (26118.64) \end{array}$	
TRI Air Lead Emissions w/in 250m in Birth Year (x100)	0.10 (3.12)	$\begin{array}{c} 0.11 \\ (3.34) \end{array}$	$0.36 \\ (5.96)$	$0.29 \\ (5.34)$	
Home Built prior to 1930	$0.08 \\ (0.27)$	0.12 (0.32)	$\begin{array}{c} 0.51 \\ (0.50) \end{array}$	$0.60 \\ (0.49)$	
Born in Chicago	$0.00 \\ (0.00)$	$0.00 \\ (0.00)$	$\begin{array}{c} 0.59 \\ (0.49) \end{array}$	$0.61 \\ (0.49)$	
Previous Cases of BLL 5+ at Coordinates	$0.01 \\ (0.09)$	$0.02 \\ (0.14)$	$0.11 \\ (0.31)$	$\begin{array}{c} 0.15 \\ (0.36) \end{array}$	
Previous Cases of BLL 10+ at Coordinates	$0.00 \\ (0.06)$	$0.01 \\ (0.08)$	$0.05 \\ (0.21)$	0.07 (0.25)	
BLL 5+	${f NaN}\ ({ m NA})$	$0.03 \\ (0.17)$	${f NaN}\ ({ m NA})$	$0.06 \\ (0.24)$	
BLL 10+	${ m NaN}$ (NA)	0.01 (0.07)	${ m NaN}$ (NA)	0.01 (0.11)	

Notes: The sample includes children born in Illinois 2010-2014. Screening status is measures by 25 months of age.

	Overall	High Risk Zip Codes
N missed: BLL 5+	5819 (5696, 5940)	$\begin{array}{c} 4671 \\ (4564, 4782) \end{array}$
N missed: BLL 10+	$ \begin{array}{c} 699\\(660,\ 741)\end{array} $	$576 \\ (541, \ 613)$
IQ cost missed (million USD): BLL 5+	$ \begin{array}{c} 429 \\ (420, 438) \end{array} $	$346 \\ (339, 355)$
IQ cost per child missed (USD): BLL 5+	73656 (73176, 74149)	74200 (73700, 74700)

Table 3: Estimated Number and Costs Associated with Missed BLL 5+ in Illinois and High Risk Zip Codes

Notes: The table shows estimates from our preferred regression forest model on the number of children with $BLL \ge 5\mu g/dL$ and $BLL \ge 10\mu g/dL$ missed currently in Illinois overall and in high-risk zip codes (where all children should be screened). We also show the total and average IQ costs associated with these undetected cases of $BLL \ge 5\mu g/dL$. In parentheses, we report confidence intervals based on 1,000 simulations of BLLs for untested children.

Table 4: Correspondence between Official IDPH Zip Code Designation and Model-Based

	ID	PH
Model	Low	High
Low	0.45	0.14
High	0.14	0.26

Notes: The table shows the share of zip codes classified as high risk under the current official IDPH designation and an alternative produced by ranking zip codes in terms of overall cost of IQ losses due to lead exposure. This exercise is carried out using the 80% training sample to predict IQ losses and the 20% holdout sample to assign zip codes.

		Random	Screening Propensity		Risk Score		Distance	
Target P-tile	Measure		Bottom	Top	Bottom	Top	Bottom	Top
Panel A:	Within Zip Codes							
50 50	IQ Cost (USD)	15339 (15306, 15373)	15333 (15268, 15397)	15335 (15268, 15394)	15290 (15228, 15355)	15367 (15304, 15427)	15209 (15150, 15270)	15403 (15342, 15465)
50 50	IQ Cost if BLL 5+ (USD)	1093 (1061, 1127)	$\begin{array}{c} 1098 \\ (1038, 1163) \end{array}$	$1100 \\ (1039, 1163)$	$1033 \\ (972, 1094)$	1126 (1064, 1187)	1018 (956, 1080)	$ \begin{array}{c} 1126\\(1063,1192)\end{array} $
75 75	IQ Cost (USD)	15626 (15597, 15655)	15625 (15579, 15670)	15606 (15560, 15654)	15582 (15534, 15625)	15667 (15619, 15712)	15527 (15480, 15573)	15717 (15669, 15765)
75 75	IQ Cost if BLL 5+ (USD)	1209 (1180, 1238)	1197 (1151, 1242)	1204 (1159, 1249)	$1146 \\ (1101, 1194)$	$ 1264 \\ (1218, 1312) $	$ \begin{array}{c} 1155\\(1110,\ 1203)\end{array} $	$ \begin{array}{c} 1249\\(1199,1294)\end{array} $
90 90	IQ Cost (USD)	15716 (15689, 15743)	15731 (15693, 15768)	15702 (15664, 15740)	15693 (15656, 15727)	15739 (15702, 15775)	15663 (15624, 15699)	15779 (15740, 15817)
90 90	IQ Cost if BLL 5+ (USD)	1235 (1209, 1262)	1233 (1195, 1270)	$ \begin{array}{c} 1231 \\ (1194, 1266) \end{array} $	1196 (1161, 1235)	1279 (1242, 1317)	$ \begin{array}{c} 1213\\(1175,1251)\end{array} $	$1261 \\ (1226, 1299)$
Panel B:	Across Zip Codes							
50 50	IQ Cost (USD)	15737 (15712, 15763)	$13431 \\ (13393, 13465)$	$21243 \\ (21127, 21353)$	$\begin{array}{c} 12119 \\ (12104, 12134) \end{array}$	$\begin{array}{c} 26370 \\ (26223, 26515) \end{array}$	$17742 \\ (17655, 17828)$	15405 (15345, 15462)
50 50	IQ Cost if BLL 5+ (USD)	$1226 \\ (1201, 1252)$	214 (187, 242)	$\begin{array}{c} 4260 \\ (4140, 4382) \end{array}$	$\begin{matrix} 0 \\ (0,0) \end{matrix}$	7659 (7501, 7826)	$2336 \\ (2247, 2433)$	784 (733, 836)
75 75	IQ Cost (USD)	15737 (15712, 15763)	13463 (13436, 13487)	19384 (19309, 19453)	$12473 \\ (12461, 12484)$	$21379 \\ (21301, 21456)$	$17330 \\ (17271, 17390)$	$\frac{14639}{(14602,\ 14676)}$
75 75	IQ Cost if BLL 5+ (USD)	$1226 \\ (1201, 1252)$	196 (178, 215)	3121 (3045, 3194)	$\begin{matrix} 0 \\ (0,0) \end{matrix}$	3986 (3904, 4072)	2106 (2044, 2168)	532 (499, 562)
90 90	IQ Cost (USD)	15737 (15712, 15763)	$13701 \\ (13680, 13722)$	$17823 \\ (17775, 17868)$	$12805 \\ (12793, 12816)$	18758 (18707, 18809)	16846 (16800, 16890)	$14573 \\ (14544, 14600)$
90 90	IQ Cost if BLL $5+$ (USD)	1226 (1201, 1252)	246 (230, 263)	2237 (2189, 2286)	$\begin{matrix} 0 \\ (0,1) \end{matrix}$	2504 (2453, 2557)	1871 (1827, 1917)	565 (540, 589)

Table 5: Estimated IQ Costs Avoided under Different Targeting Rules

Notes: Simulated effect of increasing screening randomly and under different targeting rules among children born in Illinois 2010-2014. Panel A considers increasing screening rates within zip codes. Panel B considers increasing targeting statewide irrespectively of zip code. 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%. We evaluate these screening policies under different priority systems: randomly sampling among untested children, considering children whom we estimate to have high vs. low screening propensity using a LASSO-penalized logistic regression, considering children with high vs. low lead exposure risk scores as predicted by our preferred regression forest model, and considering children who are closer vs. farther away from screening providers, a factor that influences the likelihood of getting tested (Gazze, 2022). For each different policy, we report the average monetized IQ losses from lead exposure in two ways. The first includes losses from all BLLs, even those below the intervention threshold of 5µg/dL; the second sets any losses for children below this threshold to zero.

Supplementary Tables and Figures



Figure A1: Costs of Lead Exposure from IQ Loss by Blood Lead Level

Notes: The figure plots the estimated costs of having a blood lead level above 0 in terms of monetized IQ costs. We take the average IQ point loss per $1\mu g/dl$ for different levels of exposure from Lanphear et al., 2005 and Gould, 2009. This is 0.513 for BLLs $\leq 10\mu g/dl$, 0.19 for BLLs $10 - 19\mu g/dl$, and 0.11 for BLLs $\geq 20\mu g/dl$. We monetize these losses considering that one IQ point decrease for a three year old is associated with a present value earnings loss of \$20,568 in 2019 dollars (Klemick, Mason, and Sullivan, 2020). We censor the x axis at 30 for ease of visualization, but costs increase linearly after that.

Figure A2: Share of Children with BLLs 5+ by Risk Score: Continuous Regression Random Forest, Holdout Sample



Share of Children with BLLs 5+

Notes: The figure plots the proportion of children in the holdout set with $BLL \ge 5\mu g/dL$ in each vigintile (5% bin) of risk score as predicted by our winning model, the random forest continuous regression. The dashed horizontal line represents the overall share of children with BLLs 5+ in the holdout sample, just under 5%.





Histogram of Screening Rates by Zip Code

Notes: The figure plots the distribution of screening rates by 25 months of age by zip code of birth in high risk zip codes in Illinois for birth cohorts in our sample (2010-2014).

Variable	Variable (cted)
Birth Yr	High Risk Zip Code
Gender	Zip Code Neighboring High Risk Zip Code
Low Birth Weight Bins	Share Rentals in Block Group
Estimated Gestation Length	Share in Poverty in Block Group
Father Age	Share Female-Headed HHs in Block Group
Birth Weight	Median House Age in Block Group
Father Race	Block Group Population
Mother Race	Share White in Block Group
Mother Education	Share Black in Block Group
Father Education	Share Hispanic in Block Group
County Indicators	Median House Value in Block Group
Twin Birth	Share Urban in Block Group
Birth Order	Median Income in Block Group
Single Mother	Share Insured through Employer in Block Group
Apgar Score	Share Directly Purchasing Insurance in Block Group
Mother Age	Share on Medicare in Block Group
Father Hispanic	Share on Medicaid in Block Group
Child Hispanic	Share Uninsured in Block Group
Home Construction Decade	Share Pre1939 Homes in Block Group
Effective Home Construction Decade (if Renovated)	Share Pre1949 Homes in Block Group
House Is Single-family	Share Pre1979 Homes in Block Group
Primary Road within 15m	Share Post1999 Homes in Block Group
Primary Road within 30m	Month of Birth Indicators
Primary Road within 50m	Case of BLL $10+$ within a Yr of Birth and w/in $15m$
Primary Road within 100m	Case of BLL $10+$ within a Yr of Birth, $15-30m$
Primary Road within 250m	Case of BLL $10+$ within a Yr of Birth $30-50m$
Primary Road within 500m	Case of BLL $10+$ within a Yr of Birth $50-100$ m
Primary Road within 750m	Case of BLL 10+ within a Yr of Birth 100-250m
Primary Road within 1km	Case of BLL 10+ within a Yr of Birth 250-500m
Primary Road within 2km	Case of BLL $10+$ within a Yr of Birth 500-750m
TRI Air Lead Emissions 500-1000m in Birth Yr	Case of BLL 10+ within a Yr of Birth 750-1000m
TRI Air Lead Emissions 500-1000m up to Birth Yr	One Previous Case of BLL 5+ at Coordinates
TRI Water Lead Emissions 500-1000m in Birth Yr	Second Previous Case of BLL 5+ at Coordinates
TRI Water Lead Emissions 500-1000m up to Birth Yr	One Previous Case of BLL 10+ at Coordinates
TRI Soil Lead Emissions 500-1000m up to Birth Yr	Second Previous Case of BLL 10+ at Coordinates
TRI Air Lead Emissions 250-500m in Birth Yr	Share of t-1 Tract Cohort with BLLs 5+
TRI Air Lead Emissions 250-500m up to Birth Yr	Share of t-1 Tract Cohort with BLLs 10+
TRI Water Lead Emissions 250-500m in Birth Yr	Size of t-1 Tract Cohort
TRI Water Lead Emissions 250-500m up to Birth Yr	Share of t-2 Tract Cohort with BLLs 5+
TRI Soll Lead Emissions 250-500m up to Birth Ir	Share of t-2 Tract Conort with BLLS 10+
TRI AII Lead Emissions w/in 250m in Dirth Yr	Distance to Closest Open Provider
TDI Soil Load Emissions w/in 250m up to Dirth Yr	Distance to Closest Open Provider
TRI Soll Lead Emissions w/m 250m up to Birth Yr	Distance to Closest Open Provider W/ Capillary Screening
TRI AII Leau Emissions w/m 250m up to Dirth If TPI Water Lead Emissions w/m 250m up to Dirth Vr	Distance to Closest Open Flovider, 2 Distance to Closest Open Provider w/ Capillary Screening ^2
Born in Chicago	Distance to Closest Open i forider w/ Capitary Screening, 2

Notes: Distance variables at the end of the table are only included in the screening prediction model. Our pipeline creates indicators for each factor level, including missing values.

# Candidate Variables	Minimum # Observations in Node	CWTE
8	38	0.390
18	33	0.385
35	15	0.375
62	31	0.375
73	22	0.371
53	9	0.369
105	26	0.368
91	19	0.368
127	27	0.368
212	39	0.366
191	35	0.366
119	17	0.365
146	20	0.365
254	29	0.362
155	10	0.361
264	24	0.361
180	7	0.359
231	12	0.359
205	5	0.357
246	3	0.354

Table A2: Tuning Results: Continuous Regression Forest

Notes: The table shows the 20 pairs of grid values for the two tuning parameters mtry, the number of variables to consider in each recursive split, and min_n, the minimum number of observations per leaf, together with the CWTE achieved by the continuous random forest model at those parameters. Values are ordered from top to bottom in terms of highest CWTE achieved, with the winning parameters in the first row.

Model	CWTE
RF: BLL	0.386
Elastic Net: BLL	0.379
OLS: BLL	0.377
Lasso-Logit: 5+	0.370
Lasso-Logit: 5+, Downsampled	0.368
RF: 1-5, 5-10, 10-20, 20+	0.366
RF: 5+	0.364
Lasso-Logit: 10+, Downsampled	0.361
RF: 10+, Downsampled	0.359
RF: 5+, Downsampled	0.358
RF: 10+	0.338
Lasso-Logit: 10+	0.329

Table A3: Model Performance: Cost-Weighted Targeting Efficiency

Notes: This table shows the performance of each tuned model as measured by the cost-weighted targeting efficiency (CWTE). CWTE is bounded between zero and one, and such that bigger is better, with one indicating a "perfect" model. CWTE penalizes "mis-classification errors" using IQ costs. Given the ordinal "risk score" from each predictive model that ranks kids from highest to lowest risk, CWTE compares the model ranking to the true BLL ranking. We estimate four random forest models (RF) that predict different outcomes: a continuous BLL outcome; the probability of a child having a BLL that falls into one of four bins (1-5, 5-10, 10-20, 20+); the probability of a child having a BLL ≥ 5 and $\geq 5\mu$ /dL. We also estimate two Lasso-penalized logistic regression models to predict the two latter binary outcomes, along with a unpenalized OLS model and a penalized elastic net model, both of which have the same regressors and outcome variable: our continuous BLL measure. For both RF and Lasso-logit binary models we also estimate versions in which the dominant class is downsampled so that elevated and non-elevated BLLs are equally common in the downsampled data.

Ranking Variable	CWTE
Average BLL	0.276
Share 5+	0.264
Share 10+	0.220
Average Monetized IQ Losses	0.278

Table A4: Performance of Zip Code Benchmark Model

Notes: The table reports the CWTE of models that prioritize screening untested children by variables aggregated at the zip code level. Each row represents a different model.

Label	Relative Importance
Case of BLL 10+ within a Yr of Birth and w/in 15m $$	100.0
Share of t-2 Tract Cohort with BLLs 5+	29.2
Share of t-1 Tract Cohort with BLLs 5+	27.7
Share Pre1949 Homes in Block Group	25.3
Share Pre1939 Homes in Block Group	25.1
Birth Weight	24.4
Median Income in Block Group	23.7
Share on Medicaid in Block Group	20.6
Share Insured through Employer in Block Group	20.4
Size of t-2 Tract Cohort	20.1
Size of t-1 Tract Cohort	19.8
Father Age	19.5
Share Pre1979 Homes in Block Group	19.2
Share of t-2 Tract Cohort with BLLs 10+	18.9
Share of t-1 Tract Cohort with BLLs 10+	18.6
Median House Value in Block Group	18.2
Share Post1999 Homes in Block Group	17.8
Share in Poverty in Block Group	16.7

Table A5: Top Variables in BLL Regression Forest Prediction Model by Variable Importance

Notes: The table reports variable importance in the continuous random forest model for the top variables that together explain over 50% of the variation in BLLs. We normalize the importance of the top variable to 100, and report importance of each other variable relative to it. Variable importance is computed using ranger with the importance = 'impurity' option. Every time a particular predictor variable is used to make a recursive split in one of the regression trees that make up our random forest, this improves the in-sample predictive MSE. The impurity variable importance measure *averages* these improvements from a given variable across all trees, and then makes relative comparisons across variables. All variables are available both for children who were tested and children who were never tested.