Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning

Eric Potash, Joe Walsh
University of Chicago
epotash,jtwalsh@uchicago.edu

Joe Brew
University of Florida
joebrew@ufl.edu

Alexander Loewi
Carnegie Mellon University
aloewi@cmu.edu

Subhabrata Majumdar
University of Minnesota
majum010@umn.edu

Andrew Reece
Harvard University
reece@g.harvard.edu

Eric Rozier
University of Cincinnati
eric.rozier@uc.edu

Emile Jorgensen
Chicago Dept of Public Health
Emile.Jorgensen@
cityofchicago.org

Raed Mansour
Chicago Dept of Public Health
Raed.Mansour@cityofchicago.org

Rayid Ghani
University of Chicago
rayid@uchicago.edu

ABSTRACT

Lead poisoning is a problem of serious public health importance, affecting hundreds of thousands of children, and causing permanent health damage. A common approach to identifying lead hazards is to test all children for high blood lead levels and then remediate the homes of children with positive tests. This prevents children who will live at that address from getting lead poisoning, but is obviously not effective for children who have already been poisoned. This paper describes joint work with the Chicago Department of Public Health in which we address the problem of preventing childhood lead poisoning. We use historical data to build a model that predicts the risk that a child will be poisoned, so that an intervention can take place before that happens. Using 20 years of blood lead level tests, lead inspection records, property value assessments and American Community Survey data, our model allows inspectors to prioritize houses on an intractably long list of possible hazards and identify children who are at high risk of lead poisoning *before* they are poisoned. This work has been described by CDPH as pioneering in the use of machine learning and predictive analytics in public-health problems and has the potential to have a massive impact both in terms of health and economic outcomes for communities across the US.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health; K.4.1 [**Public Policy Issues**]: Human Safety

General Terms

Machine Learning, Social Good, Lead Poisoning, Public Health

1. INTRODUCTION AND MOTIVATION

Lead poisoning is a problem of serious public health importance, imposing enormous direct and lifelong costs on hundreds of thousands of children in the US. Although European states banned lead paint as early as 1909 [18], political forces and vested business interests delayed bans on leaded consumer products in the United States until the late 1970s. [20] Throughout most of the 20th century, cars ran on leaded gas, houses were coated with leaded paint, and factories emitted vast quantities of leaded waste products directly into the environment. Even now, lead in paint remains a significant hazard. For example, in Chicago, almost 90% of the housing stock was built before the ban. [12].

Exposure to lead has been found to be associated with premature birth and early neurological development issues such as edema, herniation, atrophy, and white-matter degeneration[11, 9]. Lead can cause vomiting, convulsions, paralysis, and death in high concentrations [13]. Elevated blood lead level has been found associated with lower IQ in children. A retrospective study by Mazumdar *et al* [19] shows that a 1 μg/dL increase in blood-lead level in a six-month-old child is associated with a decrease of 1 IQ point in average, and goes down to 2 IQ points on average at 10 years of age.

Because of the permanent damage it can inflict, lead brings significant indirect costs to the entire country. Based on its well documented effects on IQ and contributions to neuropsychiatric disorders such as ADHD, lead poisoning has been estimated to greatly lower lifetime earnings for individuals, and greatly increase the costs of crime prevention and special-education programs for the government. Lead-related child health issues conservatively cost over $40 billion annually [17]. Completely eliminating lead in the United States could indirectly save $200 billion dollars per year [21], ten times more money than would need to be spent on removal.

A common approach to identifying lead hazards is to test all children for high blood lead levels then remediate the homes of children with positive tests. This prevents children who will live at that address from getting lead poisoning, but it is not effective for children who have already been poisoned. The Chicago Department of Public Health (CDPH), while it devotes an enormous and concerted effort

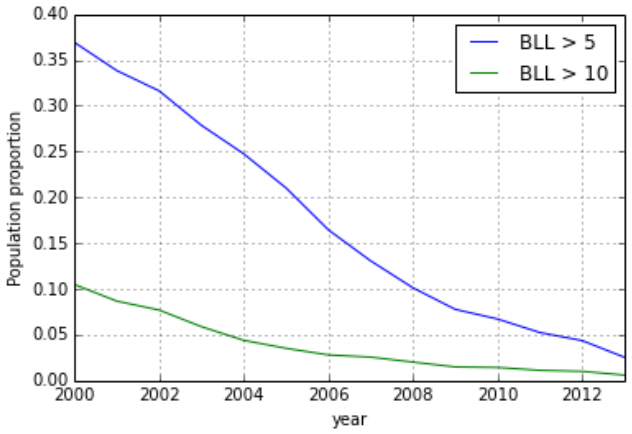

Figure 1: Blood lead tests conducted each year, blood lead tests with a concentration greater than 5 μg/dL each year, and blood lead tests with a concentration greater than 10 μg/dL each year.

to the problem of lead exposure, is typical of American cities in its damaging shortage of resources. At current levels of funding and staffing, it would take 76 years and $98 million to inspect—let alone remediate—the 197,157 old buildings in the city. The only hope of making a significant impact with the available budget is to be as smart as possible about how to use it.

This is one of the major motivations behind the work presented in this paper. In collaboration with the Chicago Department of Public Health (CDPH), we focused on identifying children who are at risk of lead poisoning and homes that are likely to have lead hazards *before* the exposure happens so the hazards can be remediated. Our approach uses historical data from 20 years of blood lead level tests, reports from home inspections for lead, and housing records to predict the risk of lead poisoning for individual children. We show that out classifier is accurate enough to identify children who are at high risk of lead poisoning *before* they are poisoned.

Based on these results, we are designing experiments to pilot the use of these predictions by CDPH to perform proactive home inspections as well as running targeted outreach ads to persuade residents living in homes at risk of lead hazards. In addition, we are working with medical providers in Chicago to implement this risk scoring into electronic medical record systems to raise early alerts for blood tests in children with high risk levels. This work has been described by CDPH as pioneering in the use of machine learning and predictive analytics in public-health problems and has the potential to have a massive impact both in terms of health and economic outcomes for communities across the US.

2. CURRENTLY USED APPROACHES

The current approaches used to deal with lead poisoning are based on testing children for lead exposure with blood tests and preventing lead exposure by inspecting homes that contain lead hazards. In this section, we describe those approaches, their shortcomings, and motivate the work we describe in this paper.

The current Center for Disease Control (CDC) recommen-

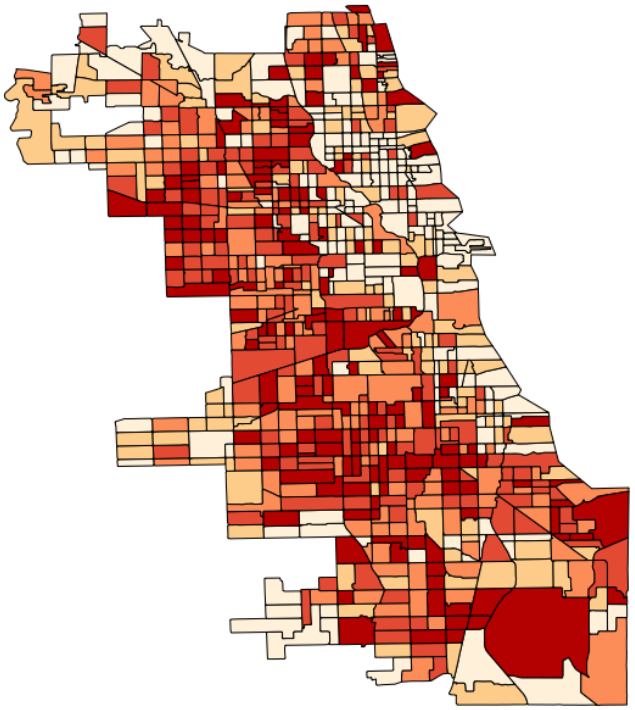

Figure 2: A heatmap of lead poisoning cases in 2012. There are spatial patterns, but space cannot adequately predict lead poisoning alone. There are too many houses in the dark-red (poisoned) areas for the city to inspect.

dations are that all children who are at risk for exposure get a lead test between 1 and 2 years of age. This is known to be the period when children start crawling and exhibiting hand-to-mouth behavior, putting them most at risk for lead dust ingestion.[16] Empirically, a sharp spike can be seen in lifetime lead levels during precisely this period, with levels peaking at 2 years of age.

Despite this being well known to public health officials, implementation of these testing recommendations is far from universal. Often, the children least at risk are the most likely to be tested early, and those most at risk often do not get tested until well after their period of greatest risk.

In addition to the CDC recommendations, many school districts, including Chicago Public Schools, require children to have had a lead test within a year prior to their matriculation. Again, this requirement may not always be adhered to and may miss the most dangerous window—the first two years of the child's life—for lead poisoning. The consequence could be many of the most vulnerable children not getting tested early enough.

Screening of youth lead levels identifies cases, but does not prevent their development. Primary prevention requires that older housing units comply with certain safety standards before they are occupied. Though screening and primary prevention work hand-in-hand, the latter is recognized professionally as more essential, and far more cost-effective, in the effort to stamp out lead poisoning. [18]

2.1 Problems with current approaches

Despite these guidelines for testing and prevention, problems remain. First, lead inspections and remediation often

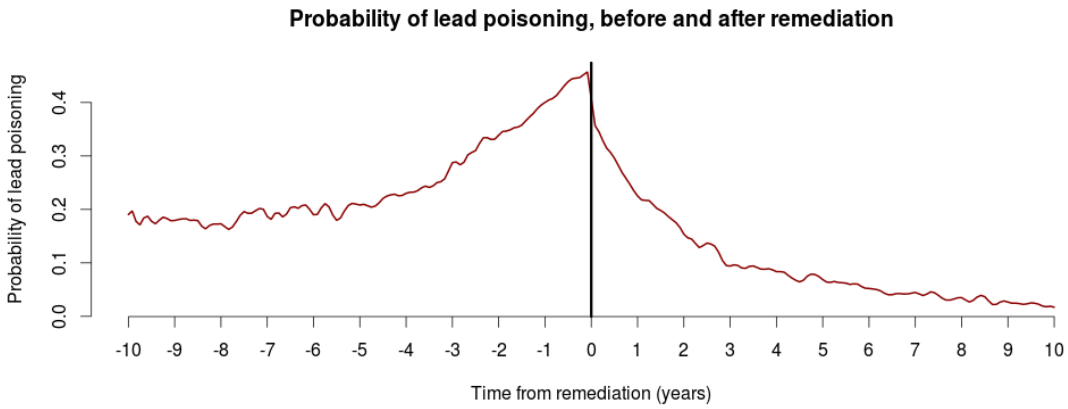

Figure 3: Mean probability of a house having $> 5\mu$g/dL BLL pre- and post-remediation

come too late, after a child has been poisoned. A positive blood test triggers and inspection, and a positive inspection triggers remediation proceedings. CDPH cannot force remediation until a child living there tests positive. Second, inspectors may focus too heavily on lead paint. While lead paint is the primary cause of poisoning [14], there are often multiple sources of poisoning, especially for higher BLLs.[10] In 5% of cases, no source can be identified [7]. Third, lead continues to poison some segments of society more heavily than others. A survey conducted in two of Chicago's riskiest neighborhoods found that 27% of children had elevated blood lead levels and 61% had never been tested previously.[12]

Secondary prevention is a challenge. Lead paint remediation (the removal of leaded paint) and soil abatement, even when conducted by a certified professional, can lead to an increase in BLLs by throwing lead particulates into the air where they are more accessible to inhalation. [8] In Chicago, most BLLs drop following the intervention (Figure 3), though the extent to which this is environmental vs. behavior-based cannot be clearly defined. Worryingly, in 9.3% of the sample, BLLs increased by at least 5 μg/dL following remediation. Approximately 46% of children who tested with BLLs that exceed 10 did not receive adequate follow-up testing [15].

2.2 Opportunities for improvement

Done well, transitioning from secondary (screening) to primary (pre-emptive remediation) prevention is an opportunity that can save resources and lives. Done poorly, it could be unnecessarily (and unfeasibly) costly, taking attention and resources away from already ill children. The fear of this latter situation is perhaps the main determinant in why most public-health agencies have not used predictive modeling for the prevention of lead poisoning. Though routine screening and secondary prevention offer a "safe" (less resource-intensive route) for policymakers and public-health practitioners, over the long-term the return on investment for primary prevention (in both health and wealth) promises to be greater.

Fortunately, the challenges of primary prevention can be mediated by the opportunities offered by recent developments in computation. Addressing both the scale and complexity of lead-related data is both feasible and affordable. Likewise, recent improvements in the quality, scope, and availability of data make the task of predictive lead poisoning prevention more feasible. The availability of public infrastructure data, combined with the digitization of medical and inspections/remediation records, offer public-health practitioners the opportunity to predict which factors cause lead poisoning, thereby enabling them to target interventions, prevent illness, and save money.

3. OUR SOLUTION

The solution we developed to predict risk of lead poisoning is based on a variety of data sources. We obtained data from the Chicago Department of Public Health that consists of blood-lead-level (BLL) test and home-inspection records, combined that with housing records and other public data (described in detail later), and built a classifier to predict the risk of childhood lead poisoning, which we define as a BLL of 5 μg/dL before the age of three years. Our system consists of the following components:

1. Data Integration and Cleaning
2. Feature Generation
3. Model Selection and Training
4. Model Validation
5. Deployment and Implementation

The next several sections describe each of the components in more detail.

4. DATA SOURCES

CDPH has two been collecting key data sources that form the basis of our predictions:

1. **Blood Lead Level Tests**: We were given the results of all 2.5 million BLL tests conducted in Chicago from 1993 through 2013. This corresponds to roughly 1 million children (see Section 5 for record linkage), with about 40,000 children born in the city every year and an average of 2.5 tests per child. Clinics submit

the BLL test results to the Illinois Department of Public Health (IDPH), and IDPH transfers the results to CDPH daily.

2. **Home Lead Inspection Records**: We were also given 120,000 home-inspection records from the same time period (1993-2013). These are detailed reports on CDPH inspector findings who were sent to a home suspected of being hazardous. The most important entries in our model are those corresponding to the date of a house's initial inspection and the date at which it was deemed to be in compliance with lead-safety standards.

We augment these two sources with a variety of publicly available data. The city's building footprint data[1] contains building characteristics such as year of construction, physical condition, number of units, stories (floors), and vacancy status. The city also provides shapefiles[1] of the census tract[2] and ward[3] boundaries. The Cook County Assessor's Office[4] has additional data on the assessed property value and building classification.

The American Community 5-year survey[6] contains census tract variables such as sociodemographics, education, health insurance, and home ownership. We also use the Census surname ethnicity data,[5] which allows estimates of ethnicity from surname alone. We combine the probability of an ethnicity given a surname with the prior probability of an ethnicity given a Census tract to get a local maximum likelihood ethnicity estimate. Ethnicity was expected to be a predictive variable because of the known and extensive history of African Americans being funneled differentially into lower-quality housing, a process known as "redlining."

5. DATA INTEGRATION AND FEATURE GENERATION

Blood test records are recorded manually and individually, so linking multiple records for a single child requires fuzzy matching of error-prone names and birthdates. We perform this using thresholded Levenshtein distances, where date of birth is a 'YYYY-MM-DD'-formatted string. Because there are millions of records, we use blocking on initials to parallelize and reduce the complexity of the computation. This process finds roughly 12% of records to contain errors in these fields.

Home addresses in the blood test records are also prone to typographic error. Roughly 20% match exactly with our address dataset. Another 75% match after cleaning using regular expressions. Another 1% are processed using a fuzzy geocoder, leaving 4% of test addresses unresolved.

Next we generate three kinds of features:

- Child features: Date of birth, imputation of ethnicity (based on census tract and last name using the census surname data), and gender (sometimes missing and sometimes conflicting between linked records). There are a total of 15 child features.

- Spatial features: After geocoding the address, we have a corresponding latitude and longitude. Using the city's shapefiles we can match this to a tract and ward (a neighborhood political boundary in Chicago). The 5-year ACS survey gives us tract-level statistics. Features include educational achievement (e.g. percentage of adults who are college graduates), income (e.g. percentage of households below the poverty line), and health (e.g. percentage of minors that are uninsured).

  The city datasets are also aggregated to the tract level, producing features such as the percentage of buildings constructed before 1978 (the year lead paint was banned), the percentage of vacant dwellings, and the average number of units per building. In total there are 44 spatial features plus indicator variables for each of Chicago's fifty wards.

- Spatio-temporal features: These are generated by aggregating the blood test and inspection records in space and time. Examples include the absolute number as well as proportion of lead poisoning cases in a Census tract in a year, the last time an address was inspected, and whether the inspection revealed interior or exterior hazards. We use a total of 29 such features.

[1]Esri vector data storage format for storing the location, shape, and attributes of geographic features

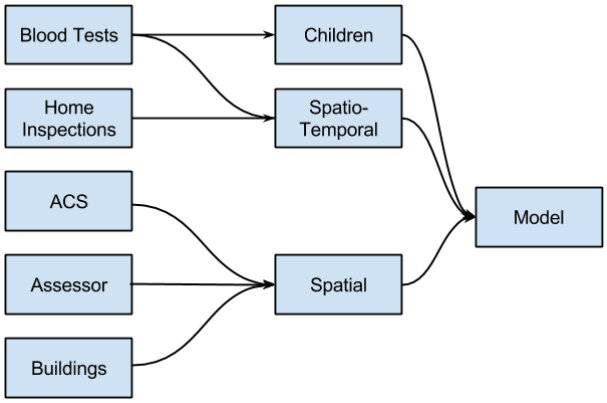

Figure 4: The data pipeline

The data pipeline is visualized in Figure 4. We used PostgreSQL with the geospatial extension PostGIS for data cleaning and aggregation. Deduplication and dataset assembly is done in Python and models are run using the scikit-learn module. The source code is available at the Data Science for Social Good GitHub repository (http://github.com/dssg).

6. EVALUATION METHODOLOGY

6.1 Cross-validation

In order to evaluate our models, we use a cross-validation strategy that emulates the way in which our models would be employed by CDPH and provides an accurate performance estimate.

For a given a point in time t_0, we train our models only on information available before t_0 and test only on information received after t_0, so that we are not training on data from the "future." The length of the training period dt is an additional necessary input to the cross validation, which we measure in years.

Our training set is thus compiled from children who received a blood test in the dt years before the t_0. For each

of these children we include a training example for the first blood test they received. If the child was below the threshold at that time but was poisoned later in the training period, we include a second training example corresponding to the first elevated blood test.

Recall that we are interested in predicting childhood lead poisoning and not individual blood samples. Thus the testing examples, unlike the training examples, correspond to children, not blood tests. A child is in the test set if they were born in the three period ending on t_0 and have not had a positive BLL test as of t_0.

There are roughly 40,000 newborns every year in Chicago, so the test set is approximately that size and the training set is approximately dt times that number.

Note this evaluation setup dictates that we cannot use the (future) dates of a child's blood test as features in predicting whether or not those tests will be positive for lead. However we can use a feature which is the minimum of the child's age at t_0 and the mean age of blood testing in the training period. Figure 5 shows an example using the cross validation of Figure 4.

Also note that the same child may appear in both the training and test periods if he or she is below the threshold before t_0 but above it after that time. Recall that we are modeling *childhood* lead poisoning so only consider blood samples up to three years after t_0. Because we train and test our models on data through 2013, we can technically only evaluate lifetime risk of children born before January 1st, 2011.

6.2 Metrics

Our models would be employed by CDPH to rank children (or buildings) according to their risk for getting (or causing) lead poisoning.

Due to limited resources, CDPH can only investigate a subset of cases. Therefore we measure the performance of a model by computing its precision in the examples predicted to be most at risk by that model. In this way we can estimate how many cases of future lead poisoning would be found, and so potentially ameliorated or avoided, if CDPH investigated a given number of cases.

Figure 6 shows the precision at different proportions of intervention for several model types evaluated at two different years.

For simplicity the evaluations below will use precision in the top 5% as the core metric. This choice is based on our observation over hundreds of model runs that this number is representative of precision at the top. That is, if model A dominates model B in the top 1% then it very unlikely that B dominates A at 10%. Note that with about forty thousand children born per year, the top 5% amounts to two thousand cases.

After finding the right model parameters we observed similar results for logistic regression, random forest and support vector machines. See Figure 7 for a comparison of the best models of each class. Thus in the results that follow, all models are logistic regressions.

7. MODELS AND RESULTS

Once our dataset is assembled, we train a variety of classification algorithms including logistic regressions, support vector machines, and random forests.

Here we present the results of running a variety of models

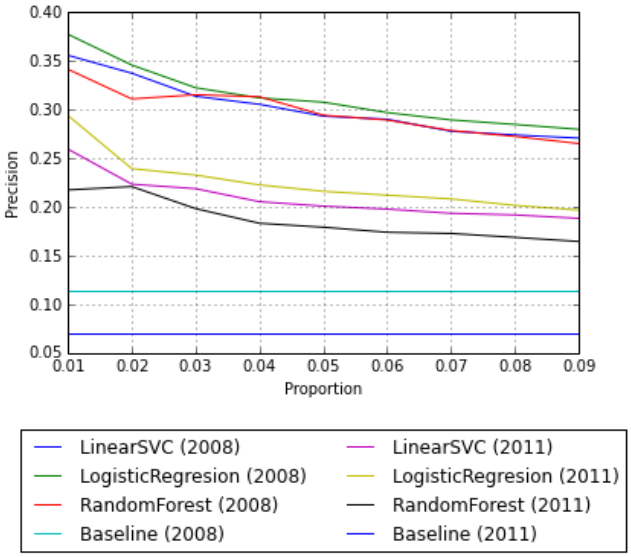

Figure 6: Precision at different proportions of investigation for different model types in different years.

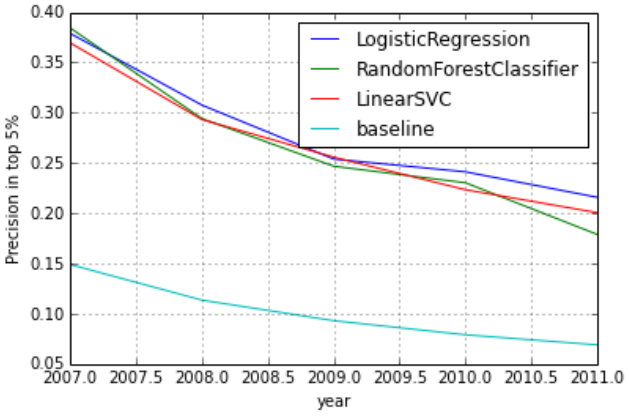

Figure 7: The best models of each class perform comparably.

Child	Birth date	Inspection date	Test date	BLL
1	1/1/2010	null	9/1/2010	1
1	1/1/2010	null	6/1/2012	7
2	6/1/2010	2/1/2011	3/1/2011	5
3	1/1/2011	1/1/2009	9/1/2011	1

→

Child	Birth date	Inspection date	Test date	y
1	1/1/2010	null	9/1/2010	False
1	1/1/2010	null	*null*	True
2	6/1/2010	*null*	*null*	False

Figure 5: Example record transformation. Changes to prevent leakage are italicized. The first row is in the training set, the next two are in the test set, and fourth row is discarded.

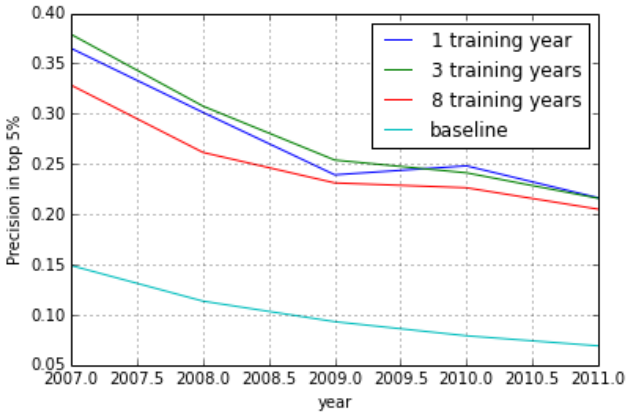

Figure 8: Performance begins to decline after about three years of training data.

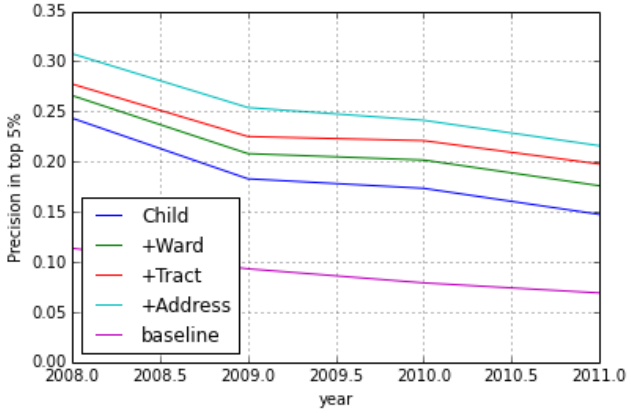

Figure 9: Precision for different feature sets

in order to find optimal parameters and measure the value of different kinds of features.

By varying the number of years in the training period, we determined that approximately three years of training data is optimal. See Figure 8. Note that the training period determines which blood samples are seen by the model but that all training examples include spatio-temporal features which draw on the entire history (blood tests and inspections) of an address.

By fitting the same model on an increasing set of features we can observe the value added by those features. Figure 9 shows that as we increase the spatial detail of our features the model improves dramatically, with address-level features (building age, condition, and history of lead poisoning and inspections) being especially important.

We can also categorize features according to the types described in Section 5. Figure 10 shows that the spatio-temporal aggregations are very important.

We use the l_1-penalized (regularization coefficient $C = .01$) logistic regression for feature selection. We examine the most important features as measured by the magnitude of their (normalized) coefficients. Figures 12 and 11 show these features having negative and positive coefficients respectively, i.e. corresponding to reduced and increased risk for lead poisoning, respectively. See the captions for a descriptions of the features.

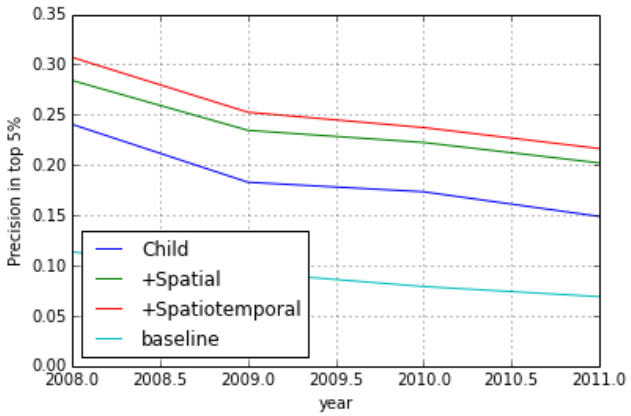

Figure 10: Increasing precision as we supplement the child features (15) with spatial features (44 plus indicators for Chicago's fifty wards) and spatio-temporal features (29).

8. IMPLEMENTATION

	c
name	
kid_birth_days	-0.932934
address_building_year	-0.198676
tract_cumulative_ebll_kid_count	-0.084628
tract_buildings_avg_year_built	-0.073164
tract_acs_edu_pct_advanced	-0.065776

Figure 11: Features with negative coefficients in a regularized logistic regression. In order these are: the child's date of birth, the year the home's construction, the proportion of children in the census tract with lead poisoning in the past year, the average year of construction, the percentage of adults in the census tract with advanced degrees.

	c
name	
tract_acs_health_pct_insured_public	0.054714
tract_acs_health_pct_uninsured	0.057941
kid_sex_M	0.064409
tract_cumulative_test_count	0.075659
address_inspection_init_days	0.080799
address_test_ebll_kid_ratio	0.169231

Figure 12: Features with positive coefficients in a regularized logistic regression. In order these are: the percentage of the population of the census tract with public health insurance, the percentage of the population without health insurance, whether the child is male, the number of BLL tests in the past year of children living in the tract, the date of the first home inspection, and the proportion of children ever poisoned at this address.

Based on the extremely encouraging results of our experiments described in the previous section, this section focuses on how CDPH is using our predictive models to prevent lead poisoning. Currently, Chicago requires doctors to determine the BLLs of all young children, regardless of housing age or the absence of other risk factors for lead exposure. This requirement is often ignored. Medicaid also requires two blood tests by age 2. In the near future, these requirements will be loosened or will be ignored even more frequently as the risk of elevated lead exposure continues its rapid decline. This will increase the need for and usefulness of a tool that allows stakeholders to better assess risk and take preventative actions, if warranted.

There are several ways that CDPH is planning to use the risk model. Each method involves a variation on disseminating the risk score to participants in a young child's life who provide medical care, child rearing, or housing and educating them on how to use this to reduce exposure to lead.

For pregnant women and parents of young children, CDPH is using billboard advertisements to encourage them to request home inspections. The risk score for these homes will be used by CDPH to prioritize inspections. In addition, publishing and publicizing the risk scores of housing allows this target audience to a) choose low risk housing when they are moving or b) request an inspection to determine if there are actual lead-based paint hazards. Even when no hazards exist, a high risk score may prompt families to make other behavior changes that minimize exposure from exterior soil (e.g. removing shoes, covering bare soil) and water (e.g. flushing) and more carefully monitor the child's diet to reduce absorption.

For doctors and other health care providers, knowing the risk score for a child can allow them to provide advice to the family regarding inspections and other exposure reducing practices. CDPH is recruiting health and social service providers to facilitate lead-based paint hazard inspections by city inspectors when their patients who are perinatal women live in high-risk housing. In addition, the CDPH is actively trying to pilot an effort where risk scores are incorporated into a child's medical record thus being available to the doctor during well-child visits.

For landlords and housing providers, CDPH is developing a program of outreach and education. For large landlords, CDPH will disseminate risk scores for their properties and encourage them to discuss and negotiate a maintenance plan with the inspectors to reduce the risk of exposure from current hazards and avoid hazardous maintenance and renovation practices. For home owners, CDPH will use the risk score to prioritize free inspections; CDPH has funding to pay for remediation for poorer owners and residents, which reduces the chance that the family will be burdened by unsustainable expenses required after the inspection.

9. THE PROMISE OF LIFETIME TRAJECTORY MODELING

While this paper describes in detail one model that was used, there are many other possible approaches to this problem and several others that have already been explored. One approach attempted to predict later lifetime exposure from infancy, using features on the child at birth, or from early doctors visits. The rationale relies on the idea that trace

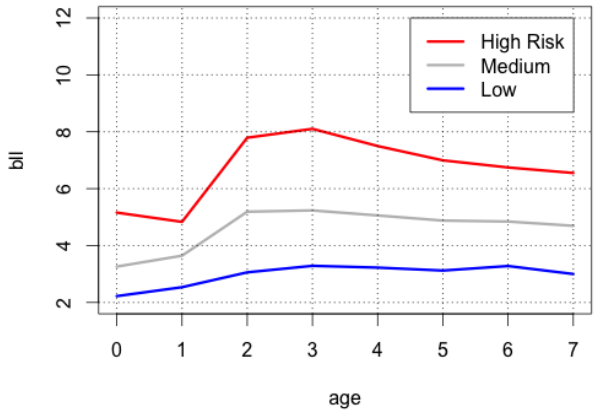

Figure 13: Lifetime trajectories from aggregated data

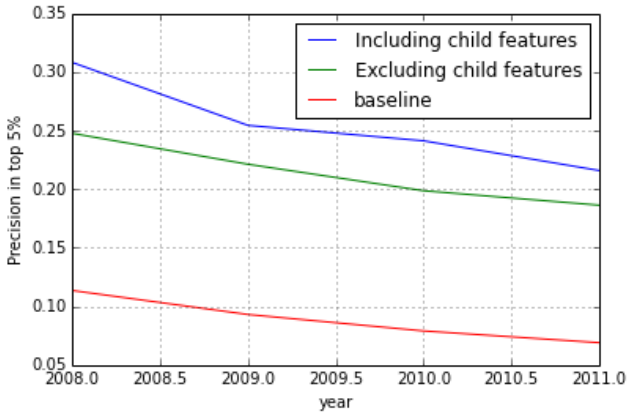

Figure 14: Precision for models with and without child features

amounts of lead, perhaps acquired from the air, or even from the mother, are indicative of much higher amounts of environmental lead. The infant is not yet at risk from these sources exclusively because of their inability to crawl. While measured levels at infancy might be below the CDC threshold, not giving a doctor quantitative cause for concern, they might also be strong indicators that the threshold will be crossed, once crawling begins – allowing a critical window of up to several months for remediation. In addition to being strictly preventative, this approach has the added significant benefit that it could be included as a part of post-natal doctors visits, meaning a schedule of opportunities for an early test is already in place.

While few features and few tests were available for an individual, there was sufficient variation in age at the time of a test to construct rough canonical lifetime trajectories by aggregating tests from many people. In other words, "average BLL at age X" could be calculated easily. One such trajectory was constructed for each census tract in the city, and these trajectories were clustered into Low, Medium, and High risk. These clusters were cleanly distinguishable even during the period prior to the jump in lead levels that takes place when crawling begins – suggesting that with more features (and aggressive early testing), this approach has significant preventative potential.

10. CONCLUSIONS AND FUTURE WORK

Thousands of Chicago children are poisoned by lead every year, incurring great health and social costs to the city in the short and long term. We developed this model in conjunction with the Chicago Department of Public Health to help them prioritize their inspection and testing schedule. Using blood tests, home inspections, county land assessments, and data from the US Census, the model produces more accurate predictions of lead risk than what CDPH had available.

CDPH is working to implement this model in several ways. CDPH has deployed billboards encouraging families to contact the city for free home lead inspections; CDPH will use the model to prioritize which houses to target first. CDPH also plans to make house-level risk scores available to the public so families can better choose where to live or, if they already live there, to minimize their risk. CDPH is also working to integrate the model into local electronic medical record systems to encourage health professionals to engage families at risk. Finally, CDPH plans to use the model to work with large landlords to rid their properties of lead hazards.

Although this model can help CDPH improve its operations, there remains room for improvement. Humans are vulnerable to lead even in the womb; we hope to get pregnancy data so that we can predict risk even before the child is born. It will also be easier for parents to intervene before the child is born.

11. ACKNOWLEDGMENTS

The work described in this paper was done partially as part of the The Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago and continued at the Center for Data Science and Public Policy. We thank our collaborators at the Chicago Department of Public Health as well as the mentors and fellows in the fellowship program for their helpful comments and feedback.

12. ADDITIONAL AUTHORS

13. REFERENCES

[1] Chicago building footprints. `https://github.com/Chicago/osd-building-footprints`.

[2] Chicago census tract boundaries. `https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Census-Tracts-2000/pt6c-hxpp`. Accessed: 2014-12-06.

[3] Chicago ward boundaries. `https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Wards/bhcv-wqkf`. Accessed: 2014-12-06.

[4] Cook county assessor's office database. `http://www.cookcountyassessor.com/`. Accessed: 2014-12-15.

[5] Demographic aspects of surnames from census 2000. `http://www.census.gov/topics/population/`

genealogy/data/2000_surnames.html. Accessed: 2014-11-25.

[6] U.S. census bureau; generated by eric potash; using american factfinder. <http://factfinder2.census.gov>. Accessed: 2015-01-05.

[7] Arizona Department of Health Services. *Annual Report 2004*. Phoenix, AZ: Bureau of Epidemiology and Disease Control, Office of Environmental Health, 2005.

[8] A. Aschengrau, A. Beiser, D. Bellinger, D. Copenhafer, and M. Weitzman. Residential lead-based-paint hazard remediation and soil lead abatement: their impact among children with mildly elevated blood lead levels. *Amer. J.Pub. Health*, 87(10):1698–1702, 1997.

[9] D. C. Bellinger. Neurological and behavioral consequences of childhood lead exposure. *PLoS Med.*, 5(5):e115, 2008.

[10] S. M. Bernard and M. A. McGeehin. Prevalence of blood lead levelsâĽě 5 μg/dl among us children 1 to 5 years of age and socioeconomic and demographic factors associated with blood of lead levels 5 to 10 μg/dl, third national health and nutrition examination survey, 1988–1994. *Pediatrics*, 112(6):1308–1313, 2003.

[11] L. M. Cleveland, M. L. Minter, K. A. Cobb, A. A. Scott, and V. F. German. Lead hazards for pregnant women and children: Part 1: Immigrants and the poor shoulder most of the burden of lead exposure in this country. part 1 of a two-part article details how exposure happens, whom it affects, and the harm it can do. *Amer. J. Nursing*, 108(10):40–49, 2008.

[12] T. A. Dignam, A. Evens, E. Eduardo, S. M. Ramirez, K. L. Caldwell, N. Kilpatrick, G. P. Noonan, W. D. Flanders, P. A. Meyer, and M. A. McGeehin. High-intensity targeted screening for elevated blood lead levels among children in 2 inner-city chicago communities. *J. Inform.*, 94(11), 2004.

[13] P. Elliott, R. Arnold, D. Barltrop, I. Thornton, I. M. House, and J. A. Henry. Clinical lead poisoning in england: an analysis of routine sources of data. *Occup.Environ. Med.*, 56(12):820–824, 1999.

[14] C. for Disease Control and Prevention. Screening young children for lead poisoning: guidance for state and local public health officials. In *Screening young children for lead poisoning: guidance for state and local public health officials*. CDC, 1997.

[15] A. R. Kemper, L. M. Cohn, K. E. Fant, K. J. Dombkowski, and S. R. Hudson. Follow-up testing among children with elevated screening blood lead levels. *J. Amer. Med. Assoc.*, 293(18):2232–2237, 2005.

[16] S. Ko, P. D. Schaefer, C. M. Vicario, and H. J. Binns. Relationships of video assessments of touching and mouthing behaviors during outdoor play in urban residential yards to parental perceptions of child behaviors and blood lead levels. *J. Exp. Sci. Environ. Epidemiol.*, 17(1):47–57, 2006.

[17] P. J. Landrigan, C. B. Schechter, J. M. Lipton, M. C. Fahs, and J. Schwartz. Environmental pollutants and disease in american children: estimates of morbidity, mortality, and costs for lead poisoning, asthma, cancer, and developmental disabilities. *Environ. Health Perspect.*, 110(7):721, 2002.

[18] B. P. Lanphear. Childhood lead poisoning prevention: Too little, too late. *J. Amer. Med. Assoc.*, 293(18):2274–2276, 2005.

[19] M. Mazumdar, D. C. Bellinger, M. Gregas, K. Abanilla, J. Bacic, and H. L. Needleman. Low-level environmental lead exposure in childhood and adult intellectual function: a follow-up study. *Environ. Health*, 10(24):1–7, 2011.

[20] H. L. Needleman. Childhood lead poisoning: the promise and abandonment of primary prevention. *Amer. J. Pub. Health*, 88(12):1871–1877, 1998.

[21] S. Zahran, H. W. Mielke, S. Weiler, and C. R. Gonzales. Nonlinear associations between blood lead in children, age of child, and quantity of soil lead in metropolitan new orleans. *Sci. Total Environ.*, 409(7):1211–1218, mar 2011.