Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning

Eric Potash
University of Chicago
epotash@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

Joe Walsh
University of Chicago
jtwalsh@uchicago.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 major public health problem that affects hundreds of thousands of children in the United States every year. A common approach to identifying lead hazards is to test all children for elevated blood lead levels and then investigate and remediate the homes of children with elevated tests. This can prevent exposure to lead of future residents, but only after a child has been poisoned. This paper describes joint work with the Chicago Department of Public Health (CDPH) in which we build a model that predicts the risk of a child to being poisoned so that an intervention can take place *before* that happens. Using two decades of blood lead level tests, home lead inspections, property value assessments, and census data, our model allows inspectors to prioritize houses on an intractably long list of potential hazards and identify children who are at the highest risk. This work has been described by CDPH as pioneering in the use of machine learning and predictive analytics in public health and has the potential to have a significant impact on both 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

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
KDD '15 , August 10-13, 2015, Sydney, NSW, Australia
©2015 ACM. ISBN 978-1-4503-3664-2/15/08 $15.00
DOI: http://dx.doi.org/10.1145/2783258.2788629.

General Terms

Machine Learning, Social Good, Lead Poisoning, Public Health, Public Policy

1. INTRODUCTION

Lead poisoning is a major public health issue, imposing lifelong health and economic costs on hundreds of thousands of children every year in the United States. Although European states banned lead paint as early as 1909 [19], political forces and vested business interests delayed bans on leaded consumer products in the United States until the late 1970s [21]. Throughout most of the 20th century, cars ran on leaded gas, houses were coated with leaded paint, and industry emitted leaded waste products directly into the environment. To this day, lead in paint remains a significant hazard. In Chicago, almost 90% of the housing stock was built before the ban [13].

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 [12, 10]. Lead can cause vomiting; convulsions; paralysis; and, in high concentrations, death [14]. Elevated blood lead levels are associated with lower IQs in children. A retrospective study by Mazumdar *et al* [20] shows that, on average, a 1 μg/dL increase in blood-lead level is associated with a decrease of 1 IQ point among six-month-olds and 2 IQ points among 10 year olds.

Because of the permanent damage it can inflict, lead poisoning imposes significant indirect costs on society at large. Based on its well-documented effects on IQ and contributions to neuropsychiatric disorders such as ADHD, lead poisoning has been estimated to significantly 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 [18]. Completely eliminating lead in the United States could indirectly save $200 billion dol-

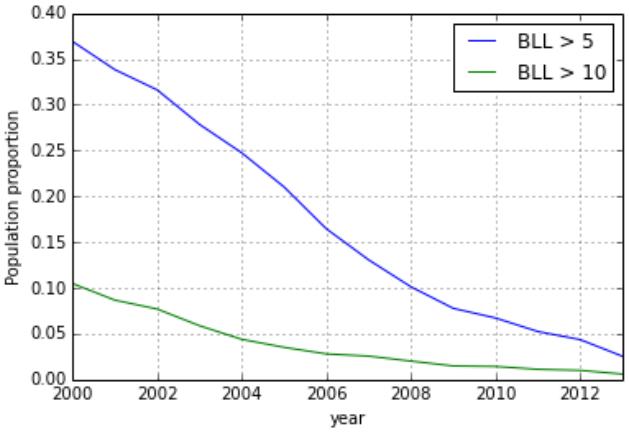

Figure 1: The proportion of blood lead tests conducted each year with a concentration greater than 5 μg/dL (blue) and greater than 10 μg/dL each year (green).

lars per year [22], ten times more money than would need to be spent on removal.

While the Chicago Department of Public Health (CDPH) devotes an enormous and concerted effort to solving the problem of lead exposure, it is like many American departments of public health in its shortage of resources. At current levels of funding and staffing, it would take CDPH 76 years and $98 million to inspect—let alone remediate—the 197,157 older buildings in Chicago. The only hope of making a significant impact with the available budget is to use it efficiently.

In collaboration with CDPH, we focused on identifying children who are at risk of lead poisoning and homes that are likely to contain lead hazards so that the hazards can be remediated *before* the children are poisoned. Our approach uses two decades of blood lead level tests, home lead inspections and remediations, housing records, and census data to build a model that can successfully predict the risk of lead poisoning for individual children.

Based on these results, we are designing experiments to pilot the use of these predictions by CDPH to perform proactive home inspections and targeted education and outreach about lead hazards. In addition, we are working with medical providers in Chicago to deploy our risk score 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 and has the potential to have a significant impact on both health and economic outcomes for communities across the US.

2. CURRENT APPROACHES

The current approaches used to deal with lead poisoning are centered around 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 the motivation for our work.

The Center for Disease Control (CDC) recommends all children at risk for exposure get a lead test between one and two 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 [17]. Our data show that lead levels rise during this period, peaking around age two.

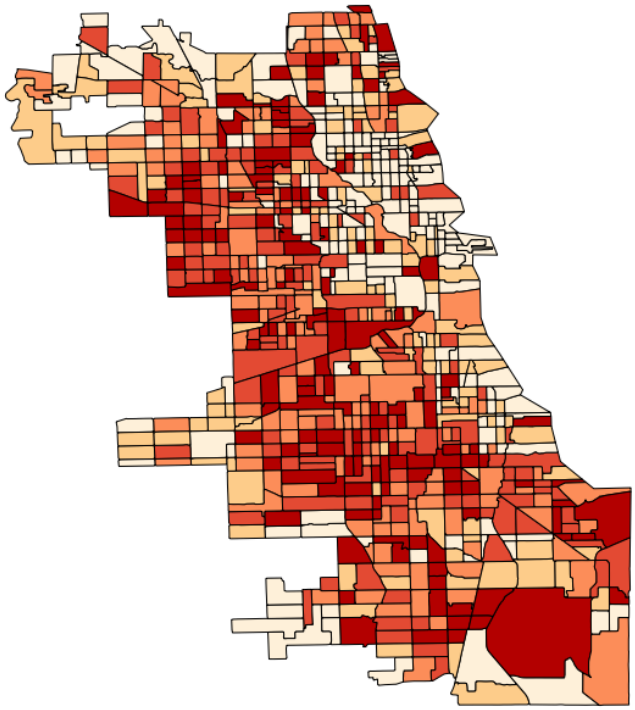

Figure 2: A heatmap of lead poisoning cases in 2012. There are spatial patterns, but space cannot adequately predict lead poisoning alone. There are too many houses in the dark-red (poisoned) areas for the city to inspect.

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 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 no more than a year before their matriculation. Again, this requirement may not always be adhered to and misses the most dangerous window—the first two years of the child's life—for lead poisoning. Many of the most vulnerable children are not getting tested early enough.

Screening of blood lead levels in children identifies cases but does not prevent their occurrence. Primary prevention requires that older housing units comply with lead safety standards before they are occupied by children. Though screening and primary prevention work are complementary, the latter is recognized to be more important and far more cost-effective in the effort to eliminate lead poisoning [19].

2.1 Problems with current approaches

Despite these guidelines for testing and prevention, problems remain. First, lead inspections and remediation often come too late, after a child has been poisoned. A positive blood test triggers an inspection, and a positive inspection triggers remediation proceedings. CDPH cannot enforce remediation until a child living in a home tests positive. Sec-

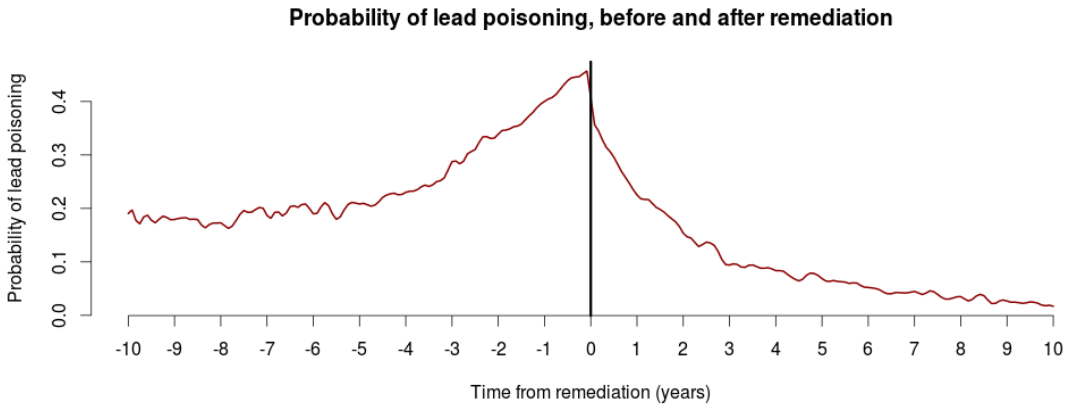

Figure 3: Mean probability of a house having $> 5\mu$g/dL BLL pre- and post-remediation

ond, inspectors may focus too heavily on lead paint. While lead paint is the primary cause of poisoning [15], there are often multiple sources of poisoning, especially for higher blood lead levels [11]. In 5% of cases, no source can be identified [8]. 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 [13].

Secondary care 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 dispersing lead particulates into the air where they are more accessible to inhalation [9]. In Chicago, most BLLs drop following the intervention (Figure 3), though the extent to which this is environmental as opposed to behavioral is unclear. Worryingly, in 9.3% of the sample, BLLs increased by at least 5 μg/dL following remediation. Approximately 46% of children with BLLs over 10 did not receive adequate follow-up testing [16].

2.2 Opportunities for improvement

Since even small quantities of lead are toxic, a transition from secondary (screening) to primary (pre-emptive remediation) care has the potential to dramatically improve health outcomes. However, if not implemented efficiently, prevention could be prohibitively expensive and cannibalize resources for remediating confirmed cases of lead poisoning. This is perhaps the chief impediment to adoption of primary care for lead poisoning.

Fortunately, these challenges of primary care can be mitigated by recent advances in computation. Addressing the scale and complexity of lead-related data is both practical and affordable. Likewise, recent improvements in the quality, scope, and availability of data make the task of predictive lead poisoning prevention feasible. The availability of public infrastructure data, combined with the digitization of medical and inspections and remediation records, offer public-health practitioners the opportunity to model the risk factors associated with lead poisoning, thereby enabling them to target interventions, prevent illness, and use their resources efficiently.

3. OUR SOLUTION

The solution we developed to predict risk for 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) tests and home-inspection records, combined that with housing records and other public data (described in detail below), and built a classifier to predict the risk of lead poisoning. The city of Chicago has adopted the CDC definition of lead poisoning of a BLL of 5 μg/dL.

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 reports detail the inspector's findings when they are 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 of the census tract [2] and ward [3] boundaries. The Cook County Assessor's Office [5] has data on the assessed property value and building classification.

The American Community Five-Year Survey [7] contains census tract variables such as socio-demographics, education, health insurance, and home ownership. We also use the census surname ethnicity data [6], 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 anticipated to be a predictive variable because of the 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 'YYYYMMDD'-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 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.

After cleaning, we collect and generate three kinds of features:

- Child features: Date of birth, imputed ethnicity (based on census tract and last name using the census surname data), and imputed gender (sometimes missing and sometimes conflicting between linked records). There are a total of 5 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-scale political boundary in Chicago).

  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. We do this spatially at the address and tract level. The temporal period is dependent on the frequency of the event (blood test or inspection) at the given spatial resolution (address or tract). Figure 4 shows the periods that we used in the final model based on exhaustive search.

  The aggregate features for inspections and tests are listed and described in Figure 5. At the spatial and temporal scales in Figure 4 there are a total of 63 such features.

  For each five year period ending in 2009 to 2013, the 5-year ACS survey gives us tract-level statistics. Features include educational achievement (e.g. percentage of adults who are college graduates), wealth (e.g. percentage of households below the poverty line), and health (e.g. percentage of minors that are uninsured). There are 21 such features.

scale	event	
	inspections	tests
address	all	3 years; all
tract	all	1 year

Figure 4: The time periods used for aggregating different events at different spatial scales.

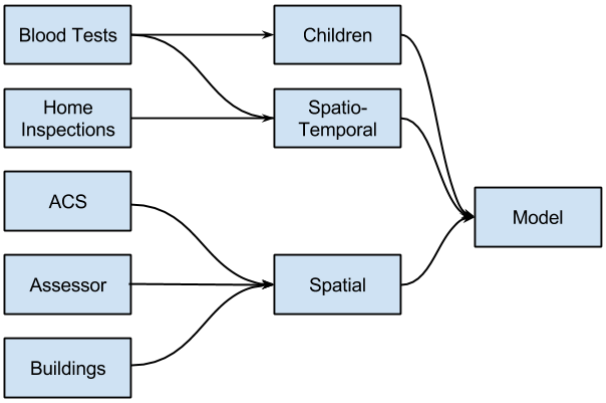

Figure 6: The data pipeline

The data pipeline is visualized in Figure 6. 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 [4].

6. EVALUATION METHODOLOGY

6.1 Cross-validation

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

For a given point in time t_0, we train our models only on information available to CDPH before t_0 to avoid 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 comprised of blood tests occurring in the dt years before the t_0. For recent times t_0, the corresponding training set contains roughly 100,000*dt examples.

feature	description
count	number of tests
tested	whether there has been a test
poisoned	whether there has been a poisoned test
ebll_count	number of poisoned tests
ebll_prop	proportion of poisoned tests
avg_bll	average blood lead level
median_bll	median blood lead level
max_bll	maximum blood lead level
min_bll	minimum blood lead level
std_bll	standard deviation of blood lead level
kid_count	number of children tested
kid_ebll_here_count	number of children with poisoned tests
kid_ebll_first_prop	proportion of children with poisoned tests
kid_ebll_first_count	number of children with first poisoned test
kid_ebll_first_prop	proportion of children with first poisoned test

(a)

feature	description
count	number of inspections
inspected	whether or not an inspection occurred
hazard_int_count	number of inspections finding interior hazards
hazard_ext_count	number of inspections finding exterior hazards
hazard_int_prop	proportion of inspections finding interior hazards
hazard_ext_prop	proportion of inspections finding exterior hazards
compliance_count	number of inspections in compliances
compliance_prop	proportion of inspections in compliance
avg_init_to_comply_days	average time from inspection to compliance

(b)

Figure 5: Spatio-temporal features for (a) blood tests and (b) home inspections.

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. Children are included in the test set if they were born before t_0 and have not been poisoned as of t_0. There are about 50,000 children in the test set for any recent t_0.

Note this evaluation setup dictates that we cannot use the (future) dates of a child's blood test as features in predicting whether those tests will return positive for lead. However, we can use the minimum of the child's age at t_0 and the mean age of blood testing in the training period. Figure 7 shows an example transformation.

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

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 8 shows the precision at different proportions of intervention for several model types evaluated at two different years. The baseline here and henceforth is given by random classification and so is equal to the incidence of lead poisoning in the test set.

For simplicity the evaluations below will use precision in the top 5% as a representative 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 the model A dominates model B at 1% then it unlikely that B dominates A at 10%. Note that 1% to 10% amounts to about 500 to 5,000 children (per year). This range is representative of the numbers that CDPH is considering for various interventions.

After tweaking each model's parameters, we observed similar performance for logistic regression, random forest, and support vector machines. See Figure 9 for a comparison of the best models of each type. Based on this observation, we use logistic regression as a representative model for the following evaluations.

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.

child	birth	inspection	test	BLL
1	2010-1-1	null	2010-9-1	1
1	2010-1-1	null	2012-6-1	7
2	2010-6-1	2011-2-1	2011-3-1	5
3	2011-3-1	2009-1-1	2011-11-1	1

(a)

child	birth	inspection	test	poisoned
1	2010-1-1	null	2010-9-1	True
1	2010-1-1	null	2012-6-1	True
2	2010-6-1	*null*	*null*	False

(b)

Figure 7: Example record transformation with t_0=1/1/2011. 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.

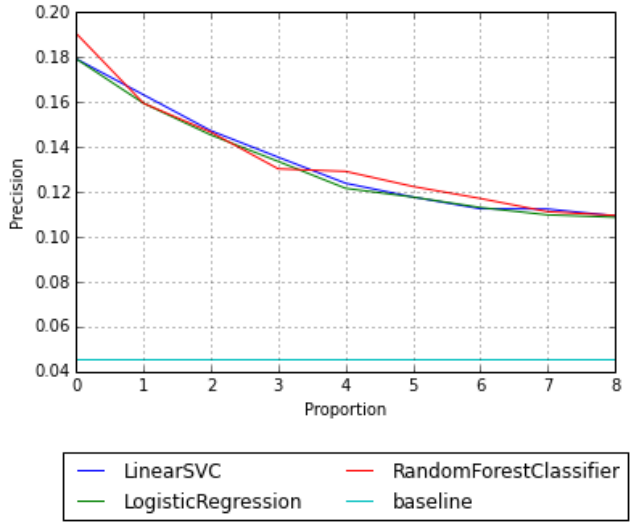

Figure 8: Precision at different proportions of investigation for different model types for the same year.

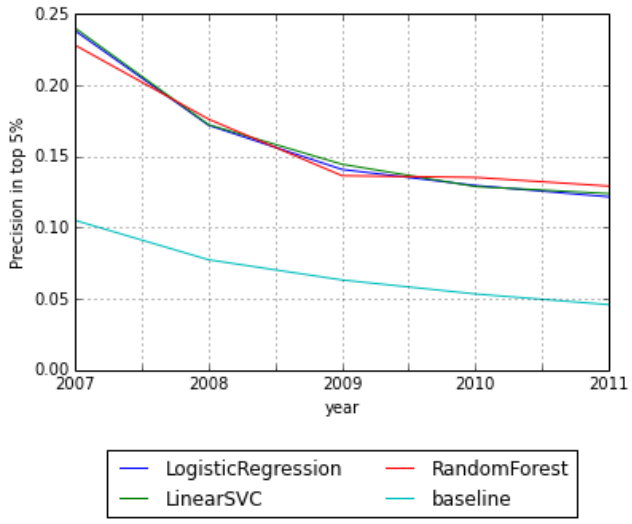

Figure 9: The best models of each class perform comparably across years.

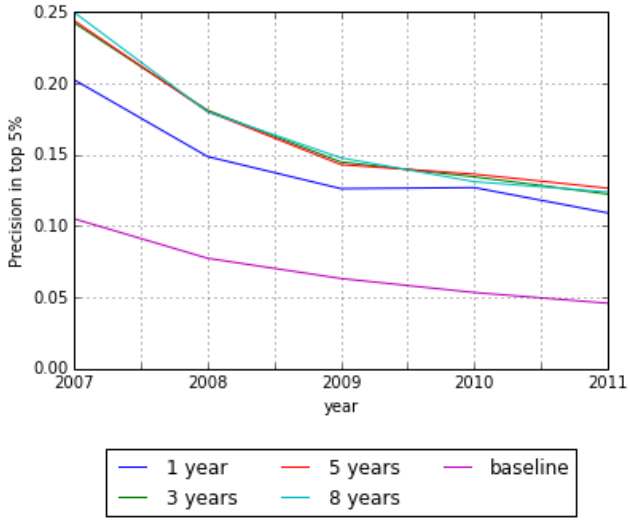

Figure 10: Performance plateaus after about three years of training data.

Here we present the results of running a variety of models 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 10. Note that the training period determines which blood samples are seen by the model but that all training examples include spatio-temporal features that draw on the entire history (blood tests and inspections) of an address or tract.

By fitting the same model on an increasing set of features we can observe the value added by those features. Figure 11 shows that as we refine the spatial scale 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 as they were presented in Section 5. Figure 12 shows that the spatial and spatio-temporal aggregations are very important.

We use the l_1-penalized (inverse regularization coefficient $C = .001$) logistic regression for feature selection. We examine the most important features as measured by the magnitude of their (normalized) coefficients. Figures 14 and 13 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.

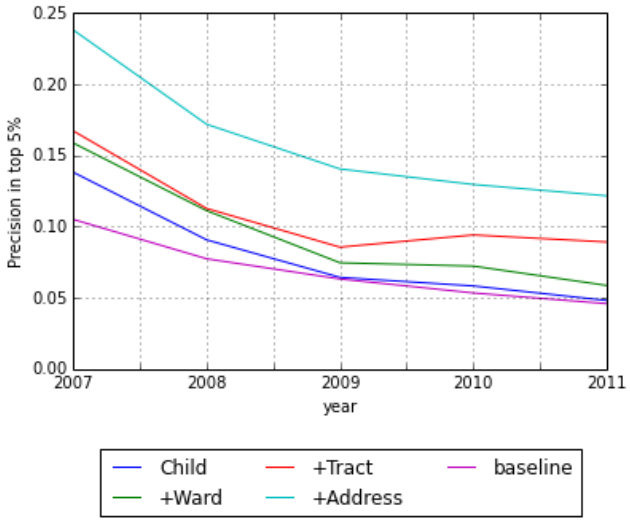

Figure 11: Precision for different feature sets

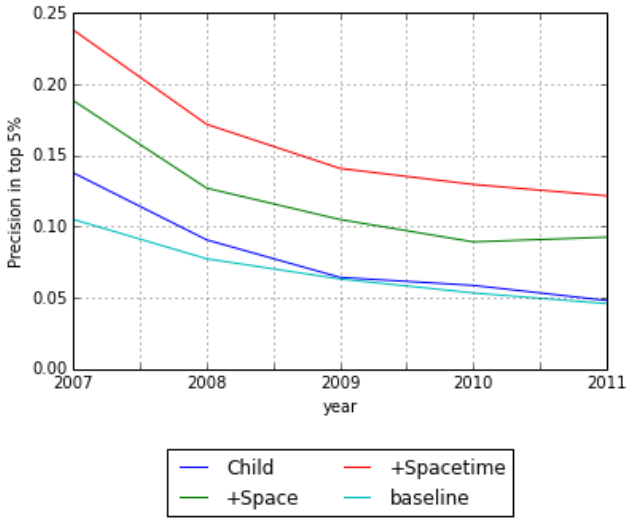

Figure 12: Increasing precision as we supplement the child features with spatial features and spatio-temporal features.

feature	coefficient
test_kid_age_days	0.224396
address_assessor_age	0.193478
address_tests_3y_kid_ebll_here_prop	0.166018
tract_tests_1y_kid_ebll_here_prop	0.145755
address_tests_3y_kid_ebll_first_prop	0.092541
address_tests_all_ebll_prop	0.052439
address_tests_3y_ebll_count	0.043917
tract_tests_1y_kid_ebll_here_count	0.028020

Figure 13: Features with positive standardized coefficients (increased risk for lead poisoning) in a regularized logistic regression. In order, these are the age of the child at the time of testing, the age of the home according to the assessor, the proportion of children tested at this address in the past three years who were poisoned first at this address, the proportion of children tested at this tract in the last year who were poisoned first at this tract, the proportion of blood tests at this address that were poisoned, the number of blood tests at this address in the past three years that were poisoned, and the number of children poisoned on this tract in the past year.

feature	coefficient
acs_5yr_race_pct_white	-0.084256
acs_5yr_health_pct_insured_employer	-0.079481
acs_5yr_edu_pct_ba	-0.020562
tract_inspections_all_compliance_prop	-0.019487
address_building_year	-0.019209
acs_5yr_edu_pct_advanced	-0.012124
address_tests_all_tested	-0.010470
address_lat	-0.007380

Figure 14: Features with negative standardized coefficients (decreased risk for lead poisoning) in a regularized logistic regression. In order, these are the percentage of the census tract that are white, the percentage of the tract that have employer health insurance, the percentage of adults with at least a bachelor's degree, the proportion of all inspections on this tract which have been complied, the year the home was built according to the buildings department, percentage of adults on the tract with an advanced degree, whether or not a child has been previously tested at this address, and the latitude of the address.

8. IMPLEMENTATION

Based on the 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 two. 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 as 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 choose low risk housing when they are moving or 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 are already being explored. One approach attempts 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 amounts of lead, perhaps acquired from the air or even from the mother, indicate 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.

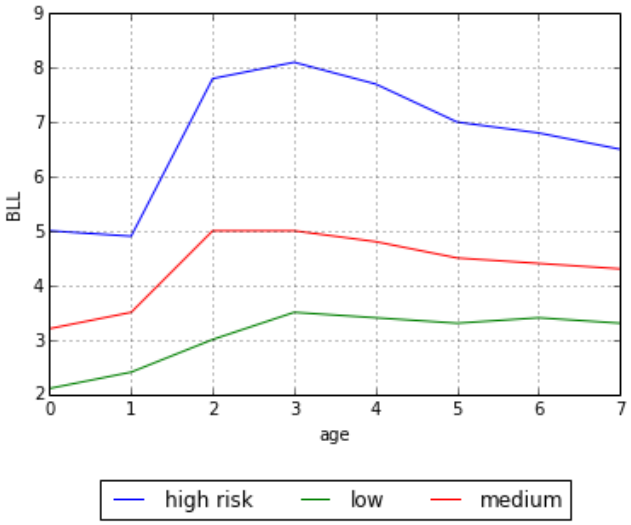

Figure 15: Lifetime trajectories from aggregated data

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 children. 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 both 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 census data, 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.

11. ACKNOWLEDGMENTS

The work described in this paper was begun 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] University of Chicago Data Science for Social Good lead model GitHub repository. `https://github.com/dssg/lead-public`. Accessed: 2015-06-09.

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

[6] Demographic aspects of surnames from census 2000. `http://www.census.gov/topics/population/genealogy/data/2000_surnames.html`. Accessed: 2014-11-25.

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

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.