Preventing Lead Poisoning in Chicago

The Eric & Wendy Schmidt **Data Science for Social Good** Summer Fellowship

Joe Brew (Univ. of Florida), Alex Loewi (Carnegie Mellon),
Subhabrata Majumdar (Univ. of Minnesota), Andrew Reece (Harvard)
Mentor: Eric Rozier (Univ. of Miami)

THE UNIVERSITY OF CHICAGO

The problem

From 1995 to 2013, 298,675 Chicago children (29% of those tested) were poisoned by lead (blood lead level [BLL] >5 micrograms per deciliter). Though the incidence of lead poisoning has declined drastically (fewer than 3% in 2013), the consequences for those sickened are severe and life-long. Lead poisoning is associated with intellectual disability, systemic organ malfunction, aggression, and in severe cases, death.

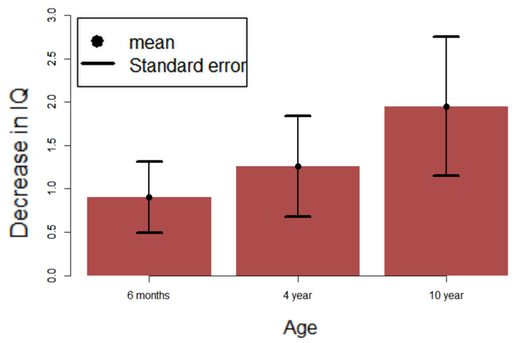


Fig. 1: Mean decrease in IQ points for 1 mcg/dL increase in BLL (source: Mazumdar et al. *Environmental Health* 2011, **10**:24)

The task

The principal pathway of childhood exposure to lead is through the home. Upon notification of a poisoned child, the Chicago Department of Public Health (CDPH) inspects that child's home for lead hazards. Our team was assigned the task of helping CDPH go from a reactive to a proactive system in which homes predicted to be a potential source of exposure would be inspected *before* a child falls ill.

Current system

Child tests positive for lead poisoning → CDPH performs a home inspection → If present, lead hazard is remediated → Child's health improves

Our goal

House is predicted to be a source of lead → CDPH performs a home inspection → If present, lead hazard is remediated → Child never gets sick

The data

We had four principal data sources:

1. 20 years of BLL tests
2. CDPH inspections data
3. Cook County assessor data
4. 2010 census

The Predictive Model

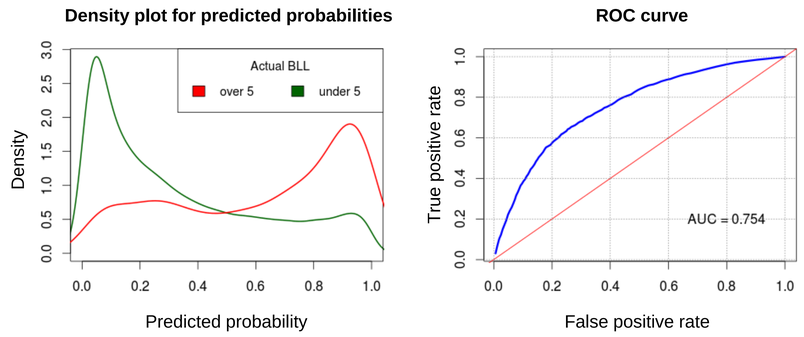


Fig 2: Model outputs: (left) Densities of two sub-populations in test data (right) ROC curve of predicted probabilities

To narrow things down, we built a suite of models for predicting BLL > 5, using features like age and condition of building, poverty rate of neighborhood, and history of lead exposure at that location. After testing, our best performing model was a downsampled aggregated random forest with the following characteristics:

- Downsampling of majority class (i.e. <5 class) to tackle rare occurrence of high BLL
- Separate random forests made on 25 rebalanced samples from training data
- All predicted probabilities for a test sample averaged to get final probability

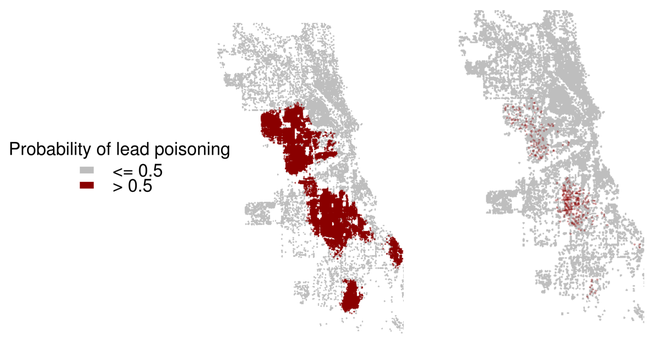


Fig. 3: Buildings targetted for inspection using current model (left) and projected model with birth data incorporated (right)

Improvements

Modeling brings the number of "high-risk" (probability of lead poisoning greater than 0.5) buildings targeted for inspection down from 200,000 to approximately 42,000. But that number still remains too high for feasibility. Our next step is to incorporate birth data so that we can target only homes with a child at the age of greatest risk (approximately 2 years old). Doing so is expected to bring our number of "high-risk" homes down to fewer than 500 (Fig. 3).

Exposure Prevention: Estimating Life Trajectories

Current medical practice is to act only once a child has exhibited a level of lead in the blood that is considered dangerous. We found canonical trajectories by imputing tract-level trajectories, then using spectral clustering to pick out examplars. These suggest a trajectory can be predicted very early in life, allowing for preventative interventions.

This work is leading us to consider policy recommendations for child lead testing that differ quite dramatically from the current practice.

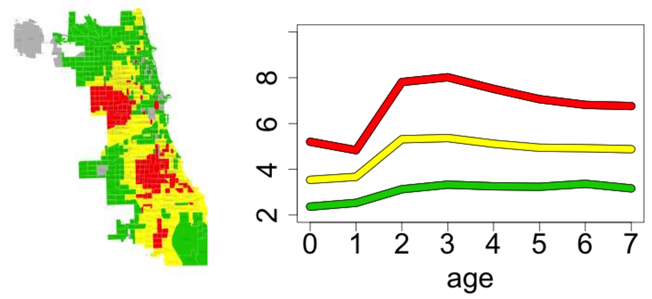
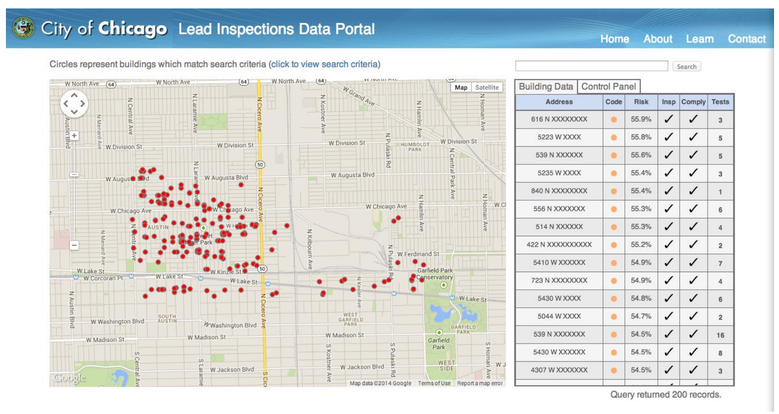


Fig. 4: Lead Trajectories (left) Risk by Census Tract (right) Mean Lead Levels at different Ages

The greatest issue with the imputation of the trajectories is the possibility of systematic bias around child age at testing. We are working with the State of Illinois to get Birth Record data to resolve this, and other, questions.

The Tool

We have built an interactive, web-based application to assist the CDPH inspections team in prioritizing which buildings to visit, as well as to understand the history of those buildings.



This work was done during the Eric & Wendy Schmidt Data Science for Social Good Fellowship at the University of Chicago.