Using Integrated City Data and Machine Learning to Identify and Intervene Early on Housing-Related Public Health Problems

Katharine Robb, DrPH, MPH; Nicolas Diaz Amigo, MPP; Ashley Marcoux, MPP; Mike McAteer, BS; Jorrit de Jong, PhD, MSc, MA

ABSTRACT

Context: Housing is more than a physical structure—it has a profound impact on health. Enforcing housing codes is a primary strategy for breaking the link between poor housing and poor health.

Objective: The objective of this study was to determine whether machine learning algorithms can identify properties with housing code violations at a higher rate than inspector-informed prioritization. We also show how city data can be used to describe the prevalence and location of housing-related health risks, which can inform public health policy and programs.

Setting: This study took place in Chelsea, Massachusetts, a demographically diverse, densely populated, low-income city near Boston.

Design: Using data from 1611 proactively inspected properties, representative of the city's housing stock, we developed machine learning models to predict the probability that a given property would have (1) any housing code violation, (2) a set of high-risk health violations, and (3) a specific violation with a high risk to health and safety (overcrowding). We generated predicted probabilities of each outcome for all residential properties in the city (N = 5989).

Results: Housing code violations were present in 54% of inspected properties, 85% of which were classified as high-risk health violations. We predict that if the city were to use integrated city data and machine learning to identify at-risk properties, it could achieve a 1.8-fold increase in the number of inspections that identify code violations as compared with current practices.

Conclusion: Given the strong connection between housing and health, reducing public health risk at more properties—without the need for additional inspection resources—represents an opportunity for significant public health gains. Integrated city data and machine learning can be used to describe the prevalence and location of housing-related health problems and make housing code enforcement more efficient, effective, and equitable in responding to public health threats.

KEY WORDS: city data, code enforcement, housing, housing inspection, innovation, machine learning

Housing is a powerful determinant of health, affecting social relationships, environmental exposures, security, and a range of other factors. Poor housing is associated with health outcomes as far reaching as cardiovascular disease, mental illness, and infectious disease.[1-4] Amid stay-at-home orders, such as during the 2020 COVID-19 pandemic, safe housing is even more critical. A nationwide study of the relationship between county-level housing conditions and COVID-19 found that poor housing (defined as overcrowding, rent >50% of income, or incomplete kitchen or bathroom facilities)

Author Affiliations: *Ash Center for Democratic Governance and Innovation, Harvard Kennedy School, Cambridge, Massachusetts (Drs Robb and de Jong, Mr Diaz Amigo, and Ms Marcoux); and Chelsea City Hall, Chelsea, Massachusetts (Mr McAteer).*

This study was funded by Bloomberg Philanthropies and the Harvard Kennedy School's Ash Center for Democratic Governance and Innovation and Rappaport Institute for Greater Boston.

The authors thank Kate Bischoff, Andy Brown, and Andrew Kieve at data analytics company, Tolemi, for their partnership in integrating city data in Chelsea. The authors thank Jonathan Jay and Elijah de la Campa for their feedback on manuscript drafts. They are grateful for the support of Chelsea City Manager, Tom Ambrosino, and the Chelsea Inspectional Services Department.

The authors declare no conflicts of interest.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (http://www.JPHMP.com).

Correspondence: *Katharine Robb, DrPH, MPH, Ash Center for Democratic Governance and Innovation, Harvard Kennedy School, 124 Mt Auburn St, Cambridge, MA 02138 (katharinerobb@hks.harvard.edu).*

DOI: 10.1097/PHH.0000000000001343

was independently associated with increased incidence and mortality from COVID-19.[5] A study in Cincinnati found that the density of housing code violations was associated with population-level morbidity independent of poverty.[6] Housing is much more than a physical structure; it has a profound impact on health.

Enforcing housing codes is a primary strategy for breaking the link between poor housing and poor health.[7] Housing codes stipulate minimum health and safety standards in rental housing. Many housing codes originated in the Sanitary Reform Movement of the late 1800s, when requirements for basic sanitation, ventilation, and other structural and hygienic conditions led to dramatic reductions in infectious disease, fire, and injury.[1,8] Modern housing inspection has transformative potential for the health of households and neighborhoods, and high-quality evidence shows the positive impact improving housing conditions has on health.[4,7] However, efforts to improve inspectional practices have not been commensurate with their critical importance.

Routine housing code enforcement falls short of its potential to effectively, efficiently, and equitably resolve housing-related health problems. It is not as *effective* as it could be because inspections are often carried out in response to residents' complaints or proactively only in limited areas. However, tenants may not report problems for fear of landlord retaliation or may be unaware that they can file complaints.[9,10] As a result, cities are often not aware of problems until they are severe, and many problems go undetected. Earlier intervention could lead to improved public health.

Inspectional practice is also not as *efficient* as it could be, because code enforcement often operates within its own department, with little coordination of data and strategies across health and other departments.[11,12] Whether a city uses a complaint-driven or a proactive approach, a lack of actionable data makes it difficult to prioritize inspections based on where limited time and resources will yield the greatest public health impact to households and neighborhoods. As a result, precious time and resources may be wasted.

Finally, code enforcement is not as *equitable* as it could be, because inspectors have broad professional discretion in the prioritization of properties for inspection. Absence of formal criteria for determining risk and need, lack of integrated data to inform that process, and the fact that the most vulnerable residents are the least likely to file a complaint increase the likelihood of unequal government protection.[9]

Given the central role of housing in health, even marginally increasing the impact of housing code enforcement offers an opportunity for significant public health gains. Investing in data analytics capabilities can make code enforcement more effective, efficient, and equitable in improving public health.[13] Cities increasingly have access to data that can be used to identify and characterize housing-related health risks and to prioritize properties for inspection and coordinated service provision. These data may come from police and fire departments, tax assessors' offices, utilities, and other sources.

Some cities already use predictive models to ensure that services and enforcement are delivered to achieve health and social goals. Washington, District of Columbia, delivers rat abatement services based on predictive modeling of high-need areas, rather than relying only on complaints.[14] In Rochester, New York, researchers used housing inspection data, alongside other data sources, to describe its potential to inform public health interventions.[15] Other studies have used machine learning and property-level data to predict which homes are at a greatest risk for vacancy.[16,17] These examples highlight how predictive analytics can inform strategic action; however, the potential of city data to identify and intervene in public health problems remains underexplored.

Objective

The objective of this study was to determine whether machine learning algorithms can outperform current practices in Chelsea, Massachusetts, in identifying properties with housing code violations that thresuaten health. We also demonstrate how integrated city data and machine learning can be used to estimate the prevalence and spatial distribution of housing-related health risks. Using administrative city data, we endeavored to predict—through machine learning models—the probability a given property would have the following: (1) any housing code violation; (2) a set of high-risk public health violations; and (3) a specific high-risk public health violation (overcrowding).

Methods

Setting

Chelsea, Massachusetts, is a small, densely populated city located just outside Boston. The majority of residents are people of color (78%), and almost half are foreign-born (46%).[18] Per capita income is $23 240 per year, making Chelsea one of the poorest cities in the state.[18] Half of the housing stock is 2- to 4-family homes, most built more than a century ago.[19] Almost 70% of residents are renters.[19] In spring 2020, COVID-19 infection rates in Chelsea were 6 times the state average.[20]

Current practices

Motivated by poor-quality housing stock, low landlord compliance, and limited housing complaints, Chelsea ran a grant-funded proactive inspection program from August 2015 to July 2018 with the goal of inspecting every rental property within a target area (N = 1263 properties inspected).[19,21] The target area was selected to encompass census block groups representative of Chelsea's rental housing and where a program existed to support low-income landlords with repairs. Violations, when found, must be resolved within a set time frame or a fine is issued. Chronic violations precipitate court action. We used data from these 1263 properties to train our models. The training data set is the best representation available of the underlying distribution of housing code violations in Chelsea because all eligible properties within the target area were inspected.

From July 2018 onward, the city continued proactive inspections but no longer restricted inspections to the target area. As of September 2019, in total 348 properties were inspected on the basis of inspector-informed prioritization. These properties, not used in model development, formed our testing data set. Under what we refer to as "current practices," inspectors meet monthly to identify blocks for inspection and track progress. Prioritization is based on inspectors' perception of the risks to residents (described in the "Results" section). Properties inspected in the last 5 years are not eligible for inspection. The testing data set allows for comparison of the machine learning results with inspector-informed prioritization and demonstrates the model performance when applied to properties outside the target area. Supplemental Digital Content Table 1 (available at http://links.lww.com/JPHMP/A766) compares properties in each data set.

Integrated city data

We worked with city staff across departments to identify and digitize administrative data sets linked to each property. These included data on housing code violations, police and fire calls, home values, and other variables (see Supplemental Digital Content Table 2, available at http://links.lww.com/JPHMP/A767). In partnership with Tolemi, a data analytics firm, and its map-based application BuildingBlocks,[22] we generated a data set consisting of each residential property in Chelsea (N = 5989) and its associated data. The change management and technical considerations associated with this process are previously published.[23]

Data cleaning

Eight of the 28 variables used in our model had missing values (1%-11% missing). We replaced missing values with the median value for numeric variables and with the most frequent occurrence in categorical variables. Incident data before September 2004 and after September 2019 were excluded.

Outcome variables

Our primary outcomes were (1) any housing code violation, (2) a set of high-risk violations, and (3) a specific high-risk violation (overcrowding). Certain serious health and housing concerns, such as lead or radon exposure, are not included in most of the housing codes and not reflected in this data set.

The "any violation" outcome indicates whether a property was issued 1 or more of 38 housing code violations used in Chelsea.[24] We selected any violation as an outcome to compare current practices with a risk-based approach using machine learning.

The "high-risk violation" outcome is a composite of any of the following 4 violations: (1) lack of or nonfunctioning smoke detectors or carbon monoxide alarms; (2) keyed locks on internal room doors, an indicator of overcrowded conditions in Chelsea; (3) infestations of insects, rodents, or skunks; and (4) accumulation of garbage in living areas. We selected high-risk violations as an outcome because not all violations represent a significant threat to public health and cities may choose to prioritize identification of higher-risk violations. Furthermore, the location and prevalence of high-risk properties are relevant not only for cities but also for other organizations that are coordinating responses (eg, health systems, community organizations). To select the high-risk violations, we began with 12 violations identified through literature review as posing an elevated health risk.[7,24] We then reviewed the list with inspectors in Chelsea and narrowed the list to 4 based on the local definition, occurrence, and severity. See Supplemental Digital Content Table 3 (available at http://links.lww.com/JPHMP/A768) for descriptions.

Finally, we selected a specific code violation from the high-risk subset: keyed locks on internal room doors. Keyed locks are placed when a home is subdivided into sublet rooms and is a strong indicator of overcrowded conditions in Chelsea.[25] We selected a single violation type as an outcome to demonstrate how integrated data and machine learning can predict specific public health risks.

Data analysis

We optimized the models for a balance of sensitivity (the proportion of properties that truly have a code violation that are predicted to have a code violation) and positive predictive value (PPV, the proportion of properties predicted to have a code violation that actually do). These optimization choices were made to (1) prioritize a subset of properties with the highest probabilities of code violations for inspection and intervention, and (2) identify as many properties in the city with a violation as possible in order to estimate the prevalence and location of housing-related health threats.

Three machine learning algorithms were compared: LASSO Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). The analysis was written in Python, using the sklearn library for running supervised learning models.[26] We selected the best performing model for each outcome based on the average precision score (APS, a summary of the sensitivity-PPV curve)[27] using 5-fold cross-validation to avoid overfitting. We report the test characteristics—sensitivity, PPV, and accuracy—and APS to describe the performance of each selected model when applied to the testing data set.

Model predictions

We assigned the predicted probability of each of the 3 outcomes to every residential property and mapped the probability range using Tableau.[28] Approximately 600 proactive inspections can be conducted per year in Chelsea. It is not possible for inspectors to inspect all properties predicted to have a code violation within a single year. Therefore, when we compare current practices with a risk-based approach, we use the threshold associated with the top-ranked 600th property in the city to generate a list of the 600 riskiest properties for each outcome. Finally, we compare features of properties predicted to have each outcome with properties not predicted to have each outcome using 2-sample t test for differences in group means.

Interviews with inspectors

To describe current practices, we conducted semistructured interviews with 2 (of 4) inspectors in Chelsea, both of whom decide how inspections are prioritized. This study was deemed exempt by the Harvard University Institutional Review Board (IRB00060687).

Results

More than half of inspected properties had at least one housing code violation (54% when combining the testing and training data sets), most of which were high risk to public health (Table 1). Overcrowding was identified in more than a quarter of properties.

TABLE 1
Proportion of Properties With Observed Housing Code Violations in Testing and Training Data Sets

Outcome[a]	Code Enforcement Practice From 2015 to 2018 (n = 1263), Training Data Set	Current Code Enforcement Practice (n = 348), Testing Data Set
Any violation	**56%**	**45%**
High-risk violation	**47%**	**40%**
No smoke detectors	33%	29%
Infestation	18%	20%
Garbage in living areas	11%	12%
Overcrowding[b]	**30%**	**27%**

[a] *Outcome variables are in bold.*

[b] *The high-risk violation also comprised the overcrowding violation (represented by keyed locks on internal room doors).*

Test characteristics

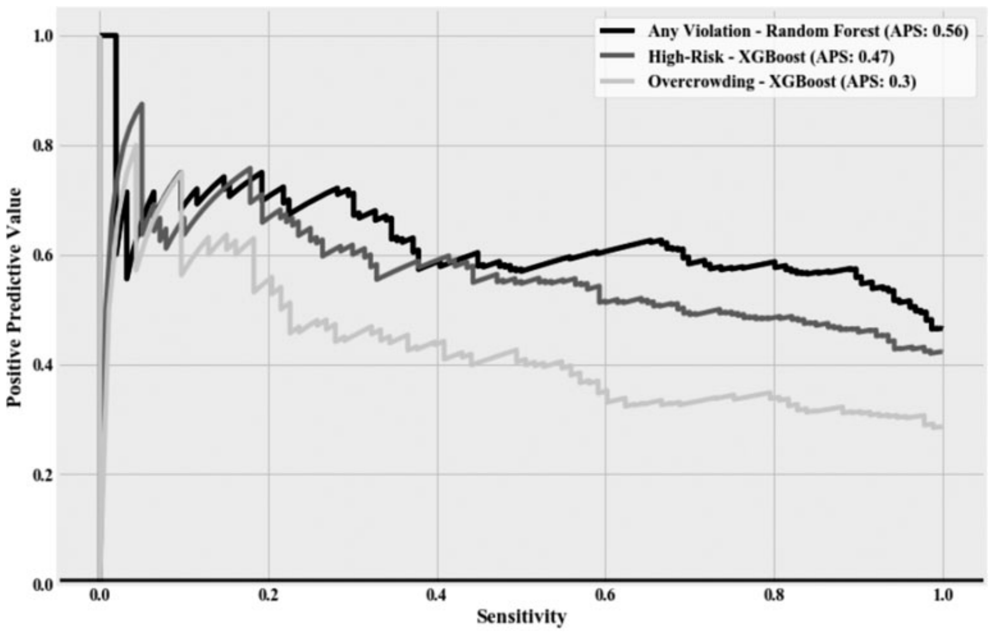
For each outcome, the best performing model was as follows: the Random Forest model predicted any violations with a sensitivity of 0.77, PPV of 0.59, and accuracy of 0.66; the XGBoost model predicted high-risk violations with a sensitivity of 0.41, PPV of 0.55, and accuracy of 0.63; and the XGBoost model predicted overcrowding with a sensitivity of 0.09, PPV of 0.62, and accuracy of 0.74. Figure 1 shows the trade-offs in sensitivity and PPV for different thresholds and the test characteristics associated with the default positivity threshold of 0.5.

Distribution and prevalence of predicted housing code violations

Maps of the estimated probabilities for each outcome reveal their spatial distribution and prevalence (Figure 2). A large portion of the city is predicted to have any code violation and high-risk violations. Predicted overcrowding is concentrated in a smaller section.

Features of properties associated with housing code violations

Properties predicted to have any of the 3 outcomes were significantly more likely to be older, have larger building size to land size ratios (indicator of housing density), and have more municipal violations (for infractions such as overgrown vegetation or uncontained trash), compared with properties predicted to not have the outcomes ($P < .001$ for all) (Table 2). See Supplemental Digital Content Table 4 (available at



Test Characteristic	Any Violation	High-Risk Violation	Overcrowding
Sensitivity	0.77	0.41	0.09
PPV	0.59	0.55	0.62
Accuracy	0.66	0.63	0.74

FIGURE 1 Tradeoffs in Sensitivity and PPV (Top) and Test Characteristics (Bottom) for Best Performing Models for Each Outcome[a]
Abbreviation: PPV, positive predictive value.
[a]See Supplemental Material Figure 1 (available at http://links.lww.com/JPHMP/A770) for the relative importance of the top 20 variables for each model.

http://links.lww.com/JPHMP/A769) for comparison across additional variables.

In interviews, housing inspectors described the most important features in identifying a property with housing code violations as (1) deteriorating conditions observed from the outside, and (2) calls from the police or fire department reporting suspected violations. They stated that for most properties there is no

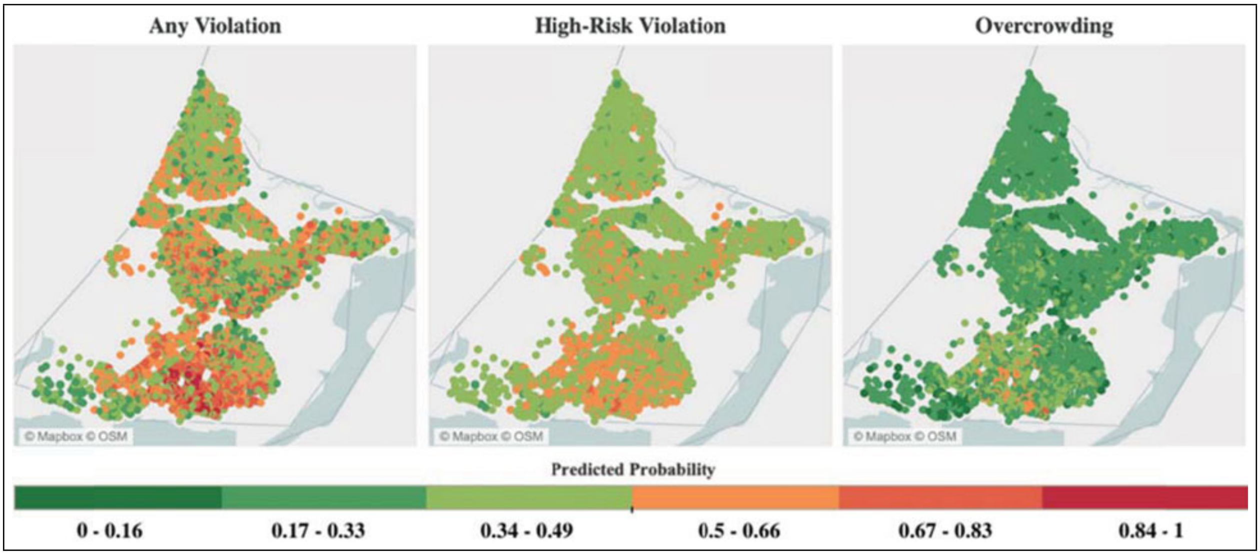


FIGURE 2 Spatial Distribution and Prevalence of Predicted Housing Code Violations in Chelsea, Massachusetts[a]
[a]Each circle represents a property and its color represents the predicted probability for each outcome. When the predicted probability is 50% or greater, properties are categorized as positive for the outcome. Circles are enlarged to protect privacy. Areas without color contain no rental properties.

TABLE 2

Differences in Properties Where Outcome Is Predicted to Be Present Versus Absent

	Any Violation, Mean (SD)	High-Risk Violation, Mean (SD)	Overcrowding, Mean (SD)
Year built (year)	Present: 1912 (25) Absent: 1935 (41) $P^a < .001$	Present: 1913 (24) Absent: 1932 (40) $P < .001$	Present: 1912 (15) Absent: 1929 (39) $P < .001$
Ratio of building size to land size[b]	Present: 1.5 (0.9) Absent: 0.6 (0.9) $P < .001$	Present: 1.7 (0.9) Absent: 0.7 (0.9) $P < .001$	Present: 2.2 (0.8) Absent: 0.9 (1.0) $P < .001$
Municipal violation (count)	Present: 8.9 (11.4) Absent: 3.5 (8.5) $P < .001$	Present: 11.1 (15.5) Absent: 3.8 (7.3) $P < .001$	Present: 15.5 (18.1) Absent: 4.9 (9.4) $P < .001$

[a] *The* P *values are calculated using a 2-sample* t *test for difference in group means.*

[b] *The city of Chelsea has a goal that ratios of land-to-building size be 1 or less to reduce high-density housing and increase green space.*

clear indication a violation will be present until an internal inspection is completed. As one inspector said, "You can have two obviously run-down properties on a street of nice homes, but you have to inspect the whole street. You'd be surprised how many of those nice homes have violations."

Prioritizing the inspection of properties with housing code violations

When applied to the entire city, our model for any violation predicts 2015 properties as having violations. If the city inspected the 600 properties (yearly capacity) with the highest probabilities of a violation, we would expect 81% to have a violation based on the model. Under current practices, 45% of inspections identify a violation. Risk-based inspection would represent a 1.8-fold increase in the number of inspections that identify code violations compared with current practices.

The same trend is observed for high-risk and overcrowding violations but with smaller increases in PPV (Table 3). Since it is unlikely a city would devote all inspections to identify a subset or single code violation, a shorter list (eg, the 100-300 highest probability properties) could be generated for these outcomes, which would also yield a higher PPV.

Discussion

Our results demonstrate how integrated city data and machine learning can (1) outperform current practices in identifying properties with code violations that threaten health, and (2) estimate the prevalence and spatial distribution of housing-related health risks.

Housing code violations in Chelsea are common, and most (85%) represent a high risk to public health. Housing codes stipulate only minimum standards for habitability; the fact that more than half of inspected properties do not meet minimum standards reveals the tip of an iceberg of housing-related health risks. Identifying and responding to housing code violations are critical public health interventions, and doing so in the most effective, efficient, and equitable manner can improve quality of life of residents and the community.

Risk-based inspection can identify at-risk properties at a higher rate than current practices, allowing intervention at a greater number of properties without the need for additional inspection resources. The

TABLE 3

Comparing Proportion of 600 Property Inspections That Result in Code Violations: Current Practices Versus Risk-Based Inspection

Outcome	Current Practices, *Observed* % (n)	Risk-Based Inspection of Top-Ranked 600 Properties, *Predicted*[a] % (n)	X-Fold Increase in Properties With Code Violations Identified
Any violation	45% (270)	81% (486)	1.8
High-risk violation	40% (240)	60% (360)	1.5
Overcrowding	27% (162)	44% (264)	1.6

[a] *Sensitivity: any violation, 16%; high-risk violation, 24%; overcrowding, 37%.*

goal of proactive inspection is prevention of violations and mitigation of risk where violations are found. Risk-based inspection prioritizes prevention and remediation efforts at the properties where inspection is estimated to have the greatest preventive or remedial effect. For example, smoke detector installation can reduce loss of life and livelihood from fire.[29] Insect extermination can reduce asthma-related emergency department visits.[30] Risk-based inspection prioritizes properties—and residents occupying them—with the greatest likelihood of need, which also promotes equity without compromising efficiency and effectiveness.

In this study, we predict that if the city were to use integrated city data and machine learning to identify at-risk properties, it could achieve a 1.8-fold increase in the number of inspections that identify code violations, as compared with current practices. One reason the models may outperform current practices is that housing inspection in Chelsea primarily targets blocks, not individual properties. Blocks include a mix of high- and low-risk properties. As inspectors stated, it is often impossible to determine which properties have violations based on outside appearance or complaints alone; therefore, they find the block-based approach practical. The block-based approach is not done for efficiency but for ease in tracking progress. The city is small (<2.5 square miles), and inspectors must contact landlords to schedule inspections. Inspection of a block area takes weeks to months, completed over many visits. A data-driven, risk-based inspection model, on the contrary, identifies specific at-risk properties anywhere in the city and targets the properties where code enforcement is likely to make the greatest difference. The models incorporate more factors into the risk calculation, allowing for detection of patterns that may not otherwise be observable.

There are few studies to compare our models with that use machine learning and city data to predict housing characteristics. One study in Cleveland used city data to predict housing vacancy. Using an XGBoost model, the authors obtained a sensitivity of 42% and a PPV of 77%; with a Random Forest model, they obtained a sensitivity of 49% and PPV of 76%,[16] both comparable with our models.

Limitations

Quality of data sets

The analysis uses city data that were not collected for research purposes. As such, the completeness and accuracy are unknown in many cases. To mitigate this limitation, we met with heads of departments to ground truth where possible. In addition, while sufficient for machine learning, the data sets were relatively small and did not allow for subanalyses.

Assumptions and biases

Chelsea does not have a rental registry, which meant that we could not separate rental units from owner-occupied units. An estimated 15% of units are owner-occupied. In addition, the model does not account for behavior change as a result of an inspection. Future studies, with better and larger data sets, can examine temporal trends and train models based on expected compliance/reinspection rates.

Finally, data-driven models contain the cognitive and social biases of the data used to create them. Biases can be actively counteracted through updating models based on use cases and engaging diverse partners in evaluating their functionality (eg, Is the model moving inspection resources to more needy parts of the city? Is it putting an unfair burden on landlords? How is it impacting tenants?).[31]

Considerations for practice

Code enforcement can improve housing conditions; however, it can also be overly punitive and lead to tenant displacement.[32] Enforcement should be coupled with service provision, where appropriate, to address root causes of housing code violations.[33,34] In 2019, Chelsea implemented a novel partnership with a local social service agency. Through this partnership, inspectors make referrals for landlords and/or tenants who face problems, such as mental illness or poverty, that make compliance with the housing code difficult.[35,36] Referrals include programs to help low-income homeowners make repairs. This gives inspectors tools beyond citations to resolve housing problems and improve public health. Outside of code enforcement, subsidy and investment in housing remain critical components of reducing public health risk.[11]

While risk-based inspection may increase efficiency, effectiveness, and equity, it should not be the only inspection method. Over time, the test characteristics of models degrade[37] and need input from new inspected properties representative of the city as it evolves, not just properties previously designated high risk. The frequency required for data source updates will depend on the pace of change in the housing stock and in population behavior.

The time, cost, and expertise needed to develop risk-based inspection models will differ widely by city, based largely on the degree of data integration and norms around the use of data for decision making. Data integration is a significant investment, but its benefits extend far beyond single initiatives or departments. While some cities have in-house capacity to develop machine learning models, others do not. If there is city-led demand, and data are integrated, the analytics work can be completed through partnerships, as was the case in this study. Many smaller cities supplement their data analytics capacity through collaboration with local universities, working with data science students and courses.[38] Models do not need to be developed de novo. Projects applying machine learning in local government often publish their code online, which can be adapted for use by others. Our code is available on GitHub.[26]

Prior to the COVID-19 pandemic, Chelsea had planned to trial risk-based inspections using results from the models; however, routine inspections were suspended as of the writing of this article. Because we did not trial the models in real life and evaluate their impact with stakeholders, we can only make inferences about their performance based on the test data.

Despite the limitations, the results of this study demonstrate the potential for increasing the public health impact of housing code enforcement through a novel application of city data.

Implications for Policy & Practice

In addition to improving housing code enforcement, using city data to estimate the prevalence and spatial distribution of housing-related health problems has important implications for policy and practice.

- Risk-based inspection can identify at-risk properties at a higher rate than inspector-informed practices alone and reduce risks to health at a greater number of properties without the need for additional inspection resources.
- Data on the prevalence and spatial distribution of housing-related risks can help cities, hospitals, and community organizations inform response activities, track improvements, and develop strategic approaches to tackle public health risk factors.[15]
- Housing data can be used to pinpoint areas of elevated risk for specific code violation(s), which can augment existing programs. For example, data on where asthma triggers are more likely can support asthma home visiting programs.
- Local data sets provide more granularity than national-level data on housing and do not require additional data collection. For research purposes, city housing data are a largely untapped data source.
- Integrated city data and machine learning are tools cities can leverage to reach more people with essential services, even on tighter budgets. During the COVID-19 pandemic, housing data and housing code enforcement as an instrument of public health are more important than ever.

Conclusion

Housing is a powerful social determinant of health. Improving housing conditions has enormous transformative potential to break the link between poor housing and poor health. Integrated city data and machine learning can be used to estimate the prevalence and spatial distribution of housing-related health problems in cities and make housing code enforcement more efficient, effective, and equitable in responding to public health threats.

References

1. Krieger J, Higgins DL. Housing and health: time again for public health action. *Am J Public Health*. 2002;92(5):758-768.
2. Adamkiewicz G, Spengler JD, Harley AE, et al. Environmental conditions in low-income urban housing: clustering and associations with self-reported health. *Am J Public Health*. 2014;104(9):1650-1656.
3. Jacobs DE. Environmental health disparities in housing. *Am J Public Health*. 2011;101(suppl 1):S115-S122.
4. World Health Organization. *WHO Housing and Health Guidelines*. Geneva, Switzerland: World Health Organization; 2018. https://www.who.int/publications-detail-redirect/9789241550376. Accessed December 17, 2020.
5. Ahmad K, Erqou S, Shah N, et al. Association of poor housing conditions with COVID-19 incidence and mortality across US counties. *medRxiv*. 2020. doi:10.1101/2020.05.28.20116087.
6. Beck AF, Huang B, Chundur R, Kahn RS. Housing code violation density associated with emergency department and hospital use by children with asthma. *Health Aff Proj Hope*. 2014;33(11):1993-2002.
7. Benjamin GC, Vernon TM. *National Healthy Housing Standard*. Columbia, MD: National Center for Healthy Housing; 2014. https://nchh.org/tools-and-data/housing-code-tools/national-healthy-housing-standard. Accessed April 12, 2019.
8. Jacobs DE, Brown MJ, Baeder A, et al. A systematic review of housing interventions and health: introduction, methods, and summary findings. *J Public Health Manag Pract*. 2010;16(5)(suppl):S5-S10.
9. Elliott DS Jr, Quinn MA. Concentrated housing code enforcement in St. Louis. *Real Estate Econ*. 1983;11(3):344-370. http://ezp-prod1.hul.harvard.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=5859978&site=ehost-live&scope=site. Accessed June 23, 2020.
10. Stacy CP, Schilling J, Barlow S, et al. *Strategic Housing Code Enforcement and Public Health*. Washington, DC: Urban Institute; 2018. https://www.urban.org/research/publication/strategic-housing-code-enforcement-and-public-health. Accessed December 4, 2018.
11. Scally CP, Waxman E, Gourevitch R, Adeeyo S. *Emerging Strategies for Integrating Health and Housing*. Washington, DC: Urban Institute; 2017. https://www.urban.org/research/publication/emerging-strategies-integrating-health-and-housing. Accessed March 16, 2019.
12. Jacobs DE, Kelly T, Sobolewski J. Linking public health, housing, and indoor environmental policy: successes and challenges at local and federal agencies in the united states. *Environ Health Perspect*. 2007;115(6):976-982.
13. Mayne Q, De Jong J, Fernandez-Monge F. State capabilities for problem-oriented governance. *Perspect Public Manag Gov*. 2020; 3(1):33-44.
14. Weaver J. *Finding the Rat: How to Optimize Your Inspections*. Cambridge, MA: Ash Center for Democratic Governance and Innovation; 2018. https://datasmart.ash.harvard.edu/news/article/

finding-rat-how-optimize-your-inspections. Accessed March 16, 2019.

15. Korfmacher KS, Holt KD. The potential for proactive housing inspections to inform public health interventions. *J Public Health Manag Pract*. 2018;24(5):444-447.
16. Martin H, Whitaker SD, Oduro I, Johnson É, Richter FG-C, Urban AH. Predictive modeling of surveyed property conditions and vacancy. In: Proceedings of the 18th Annual International Conference on Digital Government Research: dg.o '17. New York, NY: Association for Computing Machinery; 2017:358-367.
17. Appel SU, Botti D, Jamison J, Plant L, Shyr JY, Varshney LR. Predictive analytics can facilitate proactive property vacancy policies for cities. *Technol Forecast Soc Change*. 2014;89:161-173.
18. US Census Bureau. U.S. Census Bureau QuickFacts: Chelsea City, Massachusetts. https://www.census.gov/quickfacts/fact/table/chelseacitymassachusetts/RHI125216. Published 2017. Accessed June 6, 2018.
19. Ambrosino TG. City of Chelsea comprehensive housing analysis and strategic plan. https://www.chelseama.gov/sites/chelseama/files/uploads/chelsea_housing_strategy_volume_1_final_final_final.pdf. Published 2017. Accessed March 1, 2019.
20. Waller J. Here's the latest individual Mass. city and town coronavirus data. *Boston*. https://www.boston.com/news/coronavirus/2020/09/02/latest-massachusetts-city-town-coronavirus-data. Published September 2, 2020. Accessed September 21, 2020.
21. City of Chelsea. The 5 year Certificate of Habitability Rental Housing Inspection Initiative. https://www.chelseama.gov/sites/chelseama/files/uploads/5_year_coh-eng-8-2016_0.pdf. Published 2014. Accessed March 1, 2019.
22. *Tolemi*. Home page. http://www.tolemi.com/buildingblocks. Published 2020. Accessed March 2, 2019.
23. Marcoux A. Aggregate to innovate: lessons from Chelsea, Massachusetts. *Medium*. https://harvardash.medium.com/aggregate-to-innovate-lessons-from-chelsea-massachusetts-caffeab3379d. Published December 22, 2020. Accessed December 22, 2020.
24. MA Department of Public Health. Minimum Standards of Fitness for Human Habitation (State Sanitary Code). https://www.mass.gov/files/documents/2016/07/pv/105cmr410_0.pdf. Published 2007. Accessed June 23, 2020.
25. Robb K. How cities can use housing data to predict COVID-19 hotspots: lessons from Chelsea, MA. Data-Smart City Solutions. https://datasmart.ash.harvard.edu/news/article/how-cities-can-use-housing-data-predict-covid-19-hotspots-lessons-chelsea-ma. Published 2020. Accessed June 23, 2020.
26. Diaz Amigo N. Chelsea code violations and public health. https://github.com/nsdiaz/chelsea-code-violations-and-public-health. Published 2020. Accessed December 15, 2020.
27. Brownlee J. How to use ROC curves and precision-recall curves for classification in Python. Machine Learning Mastery Web site. https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python. Published August 30, 2018. Accessed June 23, 2020.
28. Tableau Public. Home page. https://public.tableau.com/en-us/s. Published 2020. Accessed June 23, 2020.
29. Ahrens M. NFPA's "Smoke alarms in U.S. home fires." https://www.nfpa.org/News-and-Research/Data-research-and-tools/Detection-and-Signaling/Smoke-Alarms-in-US-Home-Fires. Published 2019. Accessed March 7, 2019.
30. Wang C, Abou El-Nour MM, Bennett GW. Survey of pest infestation, asthma, and allergy in low-income housing. *J Community Health*. 2008;33(1):31-39.
31. Eubanks V. Automating Inequality. New York, NY: St Martin's Press; 2018. https://us.macmillan.com/automatinginequality/virginiaeubanks/9781250074317. Accessed December 17, 2020.
32. Hartman CW, Kessler RP, Legates RT. Municipal housing code enforcement and low-income tenants. *J Am Inst Plann*. 1974;40(2):90-104.
33. ChangeLab Solutions. Under One Roof. Oakland CA: ChangeLab Solutions; 2015. https://www.changelabsolutions.org/publications/under-one-roof. Accessed December 4, 2018.
34. Spivey A. On closer inspection: learning to look at the whole home environment. *Environ Health Perspect*. 2005;113(5):A320-A323.
35. Robb K. On further inspection. https://medium.com/@HarvardAsh/on-further-inspection-21794f3c840f. Published June 14, 2020. Accessed July 14, 2020.
36. Robb K. *Further Inspection: Leveraging Housing Inspectors and City Data to Improve Public Health in Chelsea, MA* [Doctoral dissertation]. Harvard T.H. Chan School of Public Health; 2019. https://dash.harvard.edu/handle/1/40976724. Accessed February 17, 2021.
37. Tsymbal A. *The Problem of Concept Drift: Definitions and Related Work*. 2004:1-8. https://www.researchgate.net/publication/228723141_The_Problem_of_Concept_Drift_Definitions_and_Related_Work. Accessed July 16, 2020.
38. MetroLab Network. About MetroLab Network. https://metrolabnetwork.org/what-we-do. Published March 26, 2019. Accessed December 15, 2020.