Precision risk assessment for pediatric hospitalization using address-level data in Cincinnati, Ohio

Carson S. Hartlage, BS[1,2], Qing Duan, PhD[3], Erika Rasnick Manning, MS[3], Judith W. Dexheimer, PhD[1,4], Andrew F. Beck, MD, MPH[4,5,6], Cole Brokamp, PhD[3,4]

1. Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, United States; 2. Department of Biostatistics, Health Informatics and Data Sciences, University of Cincinnati College of Medicine, Cincinnati, United States; 3. Division of Biostatistics and Epidemiology, Cincinnati Children's Hospital Medical Center, Cincinnati, United States; 4. Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, United States; 5. Division of General & Community Pediatrics, Cincinnati Children's Hospital Medical Center, Cincinnati, United States; 6. Division of Hospital Medicine, Cincinnati Children's Hospital Medical Center, Cincinnati, United States

The authors have no conflicts of interest to disclose.

ABSTRACT

Introduction: Persistent disparities in child health highlight the need for clinical and public health research approaches to identify and address risks with greater spatial precision. This study linked residence- and neighborhood-specific socio-environmental data to population-wide healthcare data to characterize pediatric hospitalization risk for every residential address in Cincinnati, Ohio.

Methods: We linked hospitalization data (07/01/2016-06/30/2022) to parcel-level housing data from the Hamilton County Auditor and Cincinnati Department of Buildings & Inspections and street-range crime data from the Cincinnati Police Department. Addresses were localized to 2010 census tracts to join variables from the U.S. Census American Community Survey and Eviction Lab. Generalized random forest models estimated address-level hospitalization risk and birth-adjusted hospitalization risk, accounting for child residency using vital birth records. Model performance was assessed based on varying diagnostic thresholds; fairness was evaluated by census block-level racial demographics.

Results: We matched 81.5% of hospitalizations to residential addresses. Among 77,077 addresses, 7.4% had ≥1 hospitalization. Our model performed well (ROC-AUC: 0.98–0.99; PR-AUC: 0.65-0.72) in characterizing high-risk addresses, with housing violations, violent crime, and market total value among top features. The birth-adjusted model also showed high performance (ROC-AUC: 0.92-0.93; PR-AUC: 0.65-0.78) and moderate agreement with the hospitalization risk model (κ=0.43).

Conclusions: Our results highlight the potential of address-level modeling and multiscale data integration to build on traditional area-level analyses and advance

precision population health. Future directions include geographic expansion, stakeholder engagement, and patient-level validation. This work offers a scalable approach to precisely identifying pediatric health risks, supporting targeted clinical and policy interventions.

INTRODUCTION

Despite decades of medical advances, poor and minoritized children continue to experience adverse health outcomes at disproportionately high rates. For example, children in Hamilton County, Ohio, collectively spend ~25,000 days in the hospital annually.[1] If children from all area neighborhoods were hospitalized at the same rate as those from the most affluent quartile of neighborhoods, hospitalization days could decrease by >30%.[1] These inequities have lasting effects, as early exposure to adverse socioenvironmental conditions can negatively impact health and well-being into adulthood.[2] Moreover, hospitalizations impose burdens on patients, families, and healthcare systems.[3,4]

The integration of medical, social, and environmental data can prompt a more complete understanding of how and why child health varies between and within communities and inform better strategies to address these disparities. Traditional population health approaches often rely on area-level (e.g., ZIP Code, neighborhood, census tract) characteristics. Area-based approaches, however, can be limited by shifting and ill-defined boundaries, infrequent data updates, inherent sample-based uncertainty, and the assumption of homogeneity among all individuals within the unit of area.[5] Further, deploying interventions at a neighborhood scale can be resource intensive and impractical.

Multiscale linkages are required to utilize large amounts of extant medical, social, and environmental data available at different spatiotemporal resolutions and extents. Integrated assessments of residence- and neighborhood-based influences can enhance

public health interventions and the provision of patient care but remain uncommon in population health and clinical practice.

Addresses within the electronic health record (EHR) provide an opportunity for linkages to potentially meaningful datasets, enabling more granular studies of place-based health determinants.[6] However, addresses, including those in the EHR, can be messy and unstructured, and traditional geocoding approaches are often not precise enough for address-level matching.[6–8] As such, address-level data integration is technically challenging and may be limited by poor match rates. Moreover, relevant previous work is limited in scope and has not tackled a comprehensive population study.[9–11]

Local municipalities often make address-based datasets publicly available, including information on legal units of land such as residential properties, enabling linkage of data on property type, condition, value, age, and more. This offers significantly greater spatial precision than area-based approaches, as a single census tract may contain several hundred residential properties and addresses (Figure 1).

Increased precision can come at the cost of data sparsity; hospitalizations are rare at the individual- and address-level.[12] Machine learning models are well-suited to handle this challenge, along with the collinearity and complex interactions present in many socio-environmental datasets. This makes such models valuable for integrating diverse contextual information and translating it into actionable insights for addressing health inequities.[13] Despite the availability and granularity of address-level data, there remains a lack of comprehensive models integrating residence- and neighborhood-specific data with population-wide health outcomes that can inform targeted population health and clinical practices.

We sought to develop a machine learning model that integrates population-wide healthcare data with residence- and neighborhood-specific variables to identify addresses associated with high pediatric hospitalization risk. Our dataset included 77,077 residential addresses linked to 10,085 total pediatric hospitalizations occurring during our 6-year study period and 30 input features obtained from open data sources. Using this approach, we generated address-level risk for child hospitalization (ARCH) scores for every address in Cincinnati.

MATERIALS AND METHODS

This work was completed as part of the Responding to Identified Sociomedical Risks with Effective Unified Purpose (RISEUP) Study, approved by the Cincinnati Children's Institutional Review Board. The data we used for temporal validation was covered under a second protocol, also approved by the Cincinnati Children's Institutional Review Board.

Study setting and population

This study took place in Cincinnati, Ohio, a dense urban area that is home to Cincinnati Children's Hospital Medical Center (CCHMC), the region's only pediatric hospital, and about 65,000 children and adolescents. CCHMC handles approximately 95% of all pediatric hospitalizations in Hamilton County, and coverage is likely higher within city limits, which are fully contained within the county.

The study population was defined in October 2024 as all residential addresses in the Cincinnati Area Geographic Information System (CAGIS).[14] CAGIS supports permitting, licensing, inspections, code enforcement, planning, zoning, and asset management by disseminating information on parcels, legally recognized areas of land. Addresses

associated with parcels with land use types presumed to be non-residential (e.g., park, vacant or commercial lot) and those without a land use type were excluded.

Address-level hospitalization

We used the RISEUP cohort, which consists of individuals under 18 years of age from Hamilton County, Ohio, who were hospitalized at CCHMC between July 1, 2016, and June 30, 2022.

We extracted patient addresses from the EHR and linked them to the address-based dataset of parcel identifiers obtained from CAGIS using the addr package (0.4.0) in R.[15] The purpose of addr is to clean, parse, and match addresses to a reference dataset using a natural language processing (NLP) machine learning model. Addresses associated with parcels located outside Cincinnati's boundaries were filtered using parcel centroid latitude and longitude and the city boundaries maintained by the City of Cincinnati.[16] Addresses assumed to not represent patients' long-term residences (e.g., Hamilton County Job and Family Services, CCHMC) were excluded. For each address, we calculated our primary outcome by summing the total number of hospitalizations during the study period.

To account for potential confounding by the number of children living at each address, a secondary outcome was modeled incorporating birth records as a proxy for child residency. Birth data between July 1, 2016, and June 30, 2022, were obtained from the Ohio Bureau of Vital Statistics and matched to addresses using the method described above. Birth-adjusted hospitalization scores were calculated as the number of hospitalizations minus the number of births, normalizing scores based on the likelihood that children resided at each address. For example, the birth-adjusted model would

understand a single-family home with five hospitalizations and a large apartment building with five hospitalizations as having different magnitudes of risk. Because most addresses had zero births, we chose to subtract rather than divide birth counts to avoid input scores of Infinity (e.g., one hospitalization divided by zero births) or undefined (e.g., zero hospitalizations divided by zero births).

Input Feature Data

We harmonized 30 input features for each address from public data sources: 11 at the residence level, linked via parcels or custom buffered summaries of point- or grid-based geospatial data, and 19 at the neighborhood level, linked via census tract geographies (Table 1).

Table 1. Characteristics of 77,077 addresses included in the study. IQR: Interquartile range.

Characteristic			N (%) or Median (IQR)
Address-level outcomes	**Hospitalizations**	Any Hospitalization	5,703 (7.4%)
		Hospitalizations (among addresses with at least one)	1 (1, 2)
	Births	Any Birth	19,507 (25%)
		Births (among n = 19,507 addresses with at least one)	1 (1, 2)
	Birth-adjusted hospitalization scores	Positive score (number of hospitalizations > number of births)	2,687 (3.5%)
		Score (among addresses with a positive score)	1 (1, 2)
Residence-level features	**Property Type**	Single-Family Home	54,328 (70.5)
		Two- or Three-Family Home	10,311 (13.4)
		4-19 Unit Apartment Building	4,597 (6.0)
		20-39 Unit Apartment Building	669 (0.9)
		40+ Unit Apartment Building	3,290 (4.3)
		Condominium	2,791 (3.6)

		N/A (miscellaneous)	1,091 (1.4)
	Housing Code Violations	Any Housing Code Violation	8,490 (11.0)
		Housing Code Violations (among n = 8,490 addresses with at least one)	3 (1, 6)
	Other Parcel Data	Year Built	1929 (1911, 1953)
		Market Total Value ($100k)	177.0 (121.0, 293.0)
	Crime incidents within 200 meters	Violent Crime	48 (17, 105)
		Non-violent Crime	106 (55, 197)
Neighborhood-level features	**American Community Survey**	Fraction Houses Built Before 1970	0.42 (0.28, 0.59)
		Median Home Value ($100k)	120.0 (97.2, 186.0)
		Fraction Assisted Income	0.16 (0.083, 0.28)
		Fraction High School Education	0.90 (0.81, 0.94)
		Fraction No Health Insurance	0.060 (0.030, 0.095)
		Fraction Vacant Housing	0.11 (0.075, 0.16)
		Fraction Households with Children	0.25 (0.18, 0.34)
		Number of Children per Household	0.49 (0.36, 0.62)
		Fraction Substandard Housing Conditions	0.36 (0.27, 0.43)
		Median Rent to Income Percentage	28.9 (24.4, 32.6)
		Fraction Employment	0.94 (0.90, 0.97)
		Fraction Housing Owner-Occupied	0.44 (0.31, 0.56)
		Fraction Poverty	0.19 (0.13, 0.33)
		Median Income ($10k)	45.8 (31.4, 58.5)
		Fraction Crowding	0.0080 (0, 0.026)
		Population per square km	1,920 (1,340, 2,440)
		Fraction Families with No Spouse	0.45 (0.26, 0.62)
		Fraction Limited English Speaking	0.01 (0.00, 0.02)
	Eviction Filings	Eviction Filings	227 (103, 386)

Residence-level features

Data on housing code violations by parcel identifier were obtained from the City of Cincinnati Department of Buildings & Inspections.[17] We summed the total number of

health-related (e.g., deteriorated paint, mold, pest infestation) enforced housing code violations between July 1, 2016, and June 30, 2022, for each parcel. Housing code violations in Cincinnati are complaint-driven; there are no proactive housing inspections. Other parcel data, including year built, market total value, and property type, were retrieved from CAGIS and the Hamilton County Auditor.[14,18] These data are updated at least every three years, with the last update occurring in January 2024. Property types were operationalized as shown in Table 1. Condominiums included the labels "condominium unit", "landominium", and "condo or pud garage." Separate variables were not included for less common, miscellaneous property types (e.g., "other residential structure", "metropolitan housing authority", "mobile home / trailer park"; n=1,091 addresses).

Crime data were obtained from the Cincinnati Police Department via Open Data Cincinnati.[19] These data include daily crime incidents localized to street ranges (e.g., 32XX Burnet Ave). We tabulated the number of incidents characterized in the dataset as “violent” occurring between July 1, 2016, and June 30, 2022. We summarized this number using the crime street ranges overlapping with a 200-meter (approximately 1/8-mile) radius of the parcel centroid latitude and longitude. This radius was selected based on the size of a typical city block. Incidents labelled “property” or “other” were summarized similarly as non-violent crime incidents.

For addresses matched to multiple parcels (n=522), parcel data were summarized as follows: median of market total value, year built, violent crime incidents, and non-violent crime incidents; mean number of housing code violations. Some addresses were missing data on year built (n=465) and market total value (n=127). To eliminate outliers

due to data entry errors, market total values below the 1st percentile and above the 99th percentile were considered missing (n=1,210).

Neighborhood-level features

Addresses were localized to 2010 census tract geographies by intersecting parcel centroids with tract boundary polygons and joined to U.S. American Community Survey 2019 5-year average data. Four census tracts were missing median home value data (n=1,025).

Data on eviction filings were obtained from the Eviction Lab, which began releasing tract-level data in January 2020.[20] Monthly eviction filing counts from January 2020 through June 2024 were summed. At the start of the COVID-19 pandemic, Hamilton County suspended municipal court proceedings between March 19, 2020, and June 1, 2020, and filings dropped. Our study period includes this time, and we assumed that variation among tracts did not significantly differ during periods impacted by COVID-19. Eviction data were interpolated from 2020 to 2010 census tract geographies using census block-level population weights.

Model development

We used generalized random forest models. This approach uses an adaptive nearest-neighbor method where predictions are made based on weighted averages of outcomes from data points that are “neighbors” of a target observation.[21] These local neighborhoods are adaptively defined by the structure of bagged decision trees, which partition the feature space to prioritize predictive accuracy. Random forests dynamically tailor the size and composition of local neighborhoods for each prediction, allowing them to effectively capture complex relationships in data. Because generalized random forest

handles missingness as an attribute, we did not impute missing data. Regression forest models were fit using the grf package (2.3.2) in R with 1000 trees, a sample fraction of 0.5, minimum node size of 5, and mtry of 26.

Model evaluation and calibration

Model outputs were defined as ARCH scores. Performance metrics, including area under the receiving operator curve (ROC-AUC), area under the precision-recall curve (PR-AUC), sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV), were calculated based on classification of ranked ARCH scores at representative thresholds. For the hospitalization risk model, we assessed performance for two outcomes: addresses with at least one hospitalization and addresses with at least two hospitalizations (top 7.4% and 2.4% of addresses, respectively). For the birth-adjusted model, we assessed performance for addresses with more hospitalizations than births (top 3.5%) and the top 2.4% of addresses. We evaluated the agreement between the two models using Cohen's kappa for the top 2.4% outcome. We also used calibration plots to visualize the relationship between input and output scores.

Fairness was evaluated by assessing model performance across four subgroups defined by census block-level racial composition. Addresses were localized to census blocks by intersecting parcel centroids with 2020 census block geographies and joined to 2020 U.S. Census data. They were then stratified into quartiles based on the proportion of the population in the census block identifying as white alone. Addresses missing block-level demographic data (n=83) were excluded. Model performance fairness was evaluated after applying the optimal cutoff as defined by Youden's index.

We used two fairness metrics: equalized odds, the maximum difference in sensitivity or specificity between subgroups; and equal opportunity, the maximum difference in sensitivity between subgroups.

Temporal robustness for the hospitalization risk model was assessed using address-level hospital admissions between July 1, 2022, and June 30, 2023. We re-calibrated thresholds to match the hospitalization prevalence of the validation dataset and calculated the same performance metrics described above; we also did this analysis using raw address-level hospitalization counts from the training data instead of ARCH scores. We were unable to temporally validate the birth-adjusted model because we do not have access to updated vital records data.

Model interpretation

The most important variables were determined by calculating importance as a weighted sum of how many times trees split on each feature up to a maximum depth of 4 splits. To supplement variables of importance by examining variable interactions within the random forest models, shallow decision trees were fit to the model output scores. The trees were verified by fitting decision trees to the corresponding input scores. The trees allowed us to visualize the structure and stratify ARCH scores. We also utilized partial dependence plots to visualize the relationship between market total value and risk, stratified by property type. All model interpretations are intended to be descriptive.

RESULTS

We matched 81.5% of hospital admissions to residential addresses in Cincinnati, resulting in 10,085 hospital admission encounters linked to 5,704 unique addresses. Of

77,077 total addresses, 5% had exactly one hospitalization, and 2.4% had multiple hospitalizations. In Table 1, we characterize addresses based on our input features. Most addresses are single-family homes (70.5%), with smaller proportions of multi-family homes (13.4%) and apartment buildings (11.2%). Approximately 11% of addresses had at least one housing code violation, with a median of three violations (IQR: 1, 6) among those affected. Within a 200-meter radius of each address, the median number of violent crime incidents was 48 (IQR: 17, 105), and the median number of non-violent incidents was 106 (IQR: 55, 197).

Our hospitalization risk model produced AUC values of 0.99 and 0.98 in identifying the top 2.4% and 7.4% highest-risk addresses, respectively (Table 2). PPVs were 0.34 and 0.56, respectively (Table 2). The ten most important variables were split between residence- and neighborhood-level variables, with housing violations and violent crime as the most important (Figure 2).

Table 2. Performance metrics for both hospitalization risk models for two outcomes. ROC-AUC: area under the receiving operator curve; PR-AUC: area under the precision-recall curve; CI: confidence interval; PPV: positive predictive value; NPV: negative predictive value.

Model Outcome	**Threshold**	**ROC-AUC (95% CI)**	**PR-AUC (95% CI)**	**Sens**	**Spec**	**PPV**	**NPV**
hospitalization	Multiple hospitalizations (top 2.4%)	0.99 (0.99-0.99)	0.65 (0.63-0.68)	0.97	0.95	0.34	0.99
	At least one hospitalization (top 7.4%)	0.98 (0.98-0.98)	0.72 (0.71-0.73)	0.94	0.96	0.56	0.99
birth-adjusted hospitalization	Top 2.4%	0.93 (0.92-0.95)	0.68 (0.65-0.72)	0.89	0.97	0.26	0.99

	Hospitalizations > Births (top 3.5%)	0.92 (0.91-0.93)	0.65 (0.63-0.68)	0.85	0.96	0.42	0.99

Calibration plots assessed the agreement between risk scores and actual hospitalization risk, and visual inspection revealed approximately linear relationships (Figure S1; Figure S2). For the hospitalization risk model, the median number of hospitalizations for addresses with scores above 1 was 3 greater than addresses with scores of 0 or below. Similarly, for the birth-adjusted model, the median birth-adjusted hospitalization for addresses with scores above 0.5 was 4 greater than addresses with scores below -0.5.

Birth records were linked to 19,507 addresses, or 25% of all addresses. Among these, 42.7% (n=8,129) had multiple births during the study period. Performance of the birth-adjusted model remained robust, with AUC values of 0.93 and 0.92 for the top 2.4% and top 3.5%, respectively (Table 2). The birth adjustment resulted in different rankings of the most important features (Figure 2).

We evaluated the temporal robustness of the hospitalization risk model by linking scores to address-level hospitalizations one year after the training data, resulting in 1,534 hospital admissions at 1,050 addresses (1.4% of all addresses). Performance remained fair with an ROC-AUC of 0.75 (95% CI: [0.74, 0.77]) for detecting addresses with at least one hospitalization (Table S1). Using raw hospitalization counts from the training data to forecast in lieu of ARCH scores demonstrated a slightly lower ROC-AUC of 0.68 (95% CI: [0.66, 0.69]; Table S2).

Market total value, housing code violations, violent crime, year built, and fraction of houses built before 1970 were among the ten most important variables for both models

(Figure 2). The hospitalization and birth-adjusted hospitalization risk models differed in which addresses were classified as high-risk, indicating sensitivity to outcome definition. Assessing agreement between the two models for the top 2.4% threshold produced a Cohen's kappa of 0.427, indicating moderate agreement. The property type distribution among addresses classified as high-risk by both models resembled that of the overall population (70% single-family homes; Table 1). In contrast, only 30% of addresses characterized as high-risk by the hospitalization risk model alone were single-family homes, compared to 86% for those identified solely by the birth-adjusted model.

Decisions trees fit on ARCH scores provide visualization of the most influential features and their interactions (Figure 3; Figure 4). By splitting on key residence- and neighborhood-level variables, the trees demonstrate how distinct pathways contribute to high hospitalization risk and can support interpretation of model results. Fraction of households with children, violent crime incidents, and market total value appeared in both trees, and the top split points were housing code violations and 40+ unit apartment building for the hospitalization risk model and birth-adjusted model, respectively (Figure 3; Figure 4). Partial dependence plots of market total value by property type showed a positive relationship between value and hospitalization risk and a negative relationship between value and birth-adjusted hospitalization risk, with differences between property types (Figure S3; Figure S4).

Avondale is a neighborhood with one of the highest hospitalization rates in the city.[1] The tracts shown in Figure 1 are contained in Avondale, and tract-level proportion of addresses classified as high-risk by the hospitalization risk model at the top 7.4% threshold ranged from 11.0% to 18.4%.

Our fairness assessment demonstrated some difference in performance by census block-level racial composition. The highest metric, indicating the greatest disparity between groups, was for the birth-adjusted model at the top 2.4% outcome, where equalized odds and equal opportunity highlight that sensitivity differed by 0.14 between the first quartile (greatest proportion of white residents) and the fourth quartile (lowest proportion of white residents; Table 3). There were differences in property type distribution among quartiles; the proportion of addresses that are single-family homes in the first, second, third, and fourth quartiles was 78.1%, 70.5%, 72.1%, and 60.5%, respectively.

Table 3. Fairness metrics for hospitalization and birth-adjusted hospitalization risk models based on subgroups defined by census block-level racial demographics.

Model Outcome	Diagnostic Threshold	Equalized Odds	Equal Opportunity
hospitalization	Multiple hospitalizations (top 2.4%)	0.10	0.04
	At least 1 hospitalization (top 7.4%)	0.09	0.06
birth-adjusted hospitalization	Top 2.4%	0.14	0.14
	Hospitalizations > births (top 3.5%)	0.09	0.09

DISCUSSION

We developed machine learning models to estimate child hospitalization risk for every residential address in Cincinnati, leveraging open residence- and neighborhood-level data linked to population-wide healthcare data. We operationalized ARCH scores using ranked percentile thresholds, and our approach demonstrated strong performance in

identifying addresses at high-risk for pediatric hospitalization, both total and birth-adjusted.

While neighborhood-level studies of health risk “hot spots” are well established, the use of parcel-level data in population health research is still emerging. Previous studies have used parcel-level property value data to evaluate the impact of socioeconomic status on obesity and cancer screening.[22–24] Parcel-level lead service line data have also been used to examine racial disparities in lead exposure.[25] Several studies have leveraged parcel data to study links between housing and health. Parcel- and building-level indicators of housing quality, including housing code violations, have been linked to asthma-related emergency department use, COVID-19 infection risk, and claims data for housing-sensitive health conditions.[26–28] One study used a combination of parcel, custom location buffer, and census tract data to predict home pest and mold exposures for children with asthma.[29]

Healthcare research leveraging address-based data linkage is also limited. Geocoding approaches can result in poor match rates, so direct address matching often requires NLP methods.[6–8] Other studies have used NLP-based address matching to identify patients residing in long-term care homes or other healthcare-associated settings, to measure residential mobility from the EHR, as well as for linkage across eviction, Medicaid, and homeless shelter records.[9–11,30,31] Our group has demonstrated this approach in Hamilton County, Ohio, to estimate lead exposure risk, leveraging both parcel data and address matching.[32] We build on this work here by implementing parcel data and address matching at a population- and city-wide scale, linking EHR data across pediatric conditions.

While other clinical machine learning models have incorporated area-level geomarkers data into patient-level predictions,[13] we use residential addresses as the unit of study. This has several advantages, including the ability to isolate housing-specific factors, emphasizing residence and neighborhood characteristics over individual clinical or demographic attributes. The prominence of residence-level features in our variables of importance graphs (Figure 2) and decision trees (Figure 3; Figure 4) also underscores the value of this approach. In addition, it is privacy-preserving, as ARCH scores do not link back to any individual's protected health information (since many addresses share the same score) and demonstrate better forecasting performance than historical hospitalization counts alone. This allows us to share granular health insights with interested partners without compromising patient privacy. Finally, address-level risk identification provides a foundation for precision population health strategies. Precision population health is an emerging field that aims to deliver "the right intervention to the right population at the right time," leveraging accurate and timely data; ARCH scores are well-suited to these approaches.

ARCH scores can drive primary and secondary prevention efforts in both clinical and policy spheres. In the clinic, they could be used to inform referral to medical-legal partnerships[33] or other housing advocates and targeted screening and guidance for housing-related health conditions and environmental exposures. In the policy sphere, insights from our model could drive targeted housing inspection for code enforcement and litigation against landlords with substandard housing. Partnerships with municipalities, such as the City of Cincinnati, could leverage these insights to allocate resources more effectively in ways aimed at reducing housing-related health disparities.

Community organizations focused on neighborhood safety, housing quality, and tenant rights may also be interested in address-level health insights for advocacy and resource management. By linking address-level health insights to actionable interventions, our framework offers a comprehensive approach to mitigating risk factors and potentially reducing hospitalization rates.

We sought to account for child population distribution by modeling a birth-adjusted outcome, which normalized scores by the likelihood of children residing at each address. By using birth records, we assumed that the characteristics associated with a child being born at an address are the same as those associated with a child living at an address. This model demonstrated similar performance but changed the importance of input features and the classification of many high-risk addresses. These differences highlight that many features may reflect both the propensity for an address to be the residence of a child and the propensity for child hospitalization. The two models could be leveraged for specific use cases; for example, using ARCH scores to target addresses across the entire city for intervention and using birth-adjusted ARCH scores to assess individual-level risk for a child with a known address.

While novel, our approach has important limitations. Our study focused exclusively on Cincinnati, limiting generalizability. However, the internal validity of our models is strengthened by the near-complete population coverage (>95%) of pediatric hospitalization events and full population coverage of residential addresses. Also, while the model itself is not directly translatable to other populations or areas, we provide a framework for other researchers to adapt this approach to their population and data availability.

Although our primary goal was capturing cumulative risk rather than prediction into the future, we used one year of more recent hospital admissions data (1 year after training data) to assess the validity and temporal robustness of ARCH scores. As expected, performance metrics dropped but remained fair. ARCH scores also demonstrated better forecasting ability than previous hospitalization counts alone, highlighting the contribution of residence- and neighborhood-level features to predicting future hospitalization. Further work would be needed to investigate potential model overfitting and/or temporal drift in addresses that experience hospital admissions before using ARCH scores as a forecasting tool.

Our birth-adjusted model aimed to isolate characteristics driving high hospitalization risk by adjusting for the number of children living at each address. However, our approach had shortcomings, as property type appeared to still be an important feature in this model (Figure 2; Figure 4). Residential mobility is a major confounder of the capability of birth records to represent child population distribution. Alternative approaches could leverage EHR encounters data or commercial data sources to better estimate the number of children associated with each address over the study period.

Historical inequities in housing and healthcare make fairness a critical consideration for deployment of ARCH scores. While algorithmic fairness literature has borrowed a 0.2 threshold from legal precedent of disparate impact in employment,[34] the 0.14 difference in sensitivity seen in the birth-adjusted model for the top 2.4% outcome is still concerning. In general, model performance was lower for the quartile of addresses with the lowest census block-level white population. It is critical that ARCH scores do not contribute to further neglect of historically disinvested communities and the children who

live there. Fairness issues may be driven by several factors, including disparities in property type composition among racial demographic-based subgroups (17.6% difference between first and fourth quartiles) and bias inherent in our data sources.[35] Utilizing publicly available data, even at a high spatial resolution, does not capture dynamic, context-specific information, and parcel data lack detailed information on individual units. This is particularly salient for multi-family homes, where data validity may be lower than for one child living in a single-family home, leading to potential exposure misclassification. Similarly, using vital records to infer child residence may be associated with similar biases. For housing code violations data, complaint-driven systems have been shown to be biased, and violations may be unequally enforced across neighborhoods.[36,37] Finally, using block-level demographic data for fairness analysis has limitations, as it may not be reflective of individual residences, families, or children, and scores should be interpreted in context, with consideration of accuracy-fairness tradeoffs. Fairness mitigation approaches, including debiasing and post-calibration adjustment, should be explored in the future.

Future work should expand to include the rest of Hamilton County to allow for insights into a larger population. Expansion both within our area and to other areas around the country will require more data harmonization and increased open data efforts, and smaller municipalities may not have the technical infrastructure for data sharing. For example, crime and housing code violations data are largely unavailable in areas of Hamilton County outside of Cincinnati. In addition, greater incorporation of time-varying and historical data would improve model robustness and enable evaluation of policy impacts and evolving risk patterns, rather than relying solely on cross-sectional

snapshots to estimate average characteristics over a historical period. Further, our processes, including the use of open data, can be specialized to other geographic areas through model retraining and adaptation to the data available in other municipalities. Multi-sector collaboration and the integration of multi-dimensional data from diverse sources will be vital to the scalability and portability of efforts. For example, commercial providers such as Regrid (regrid.com) offer access to nationwide harmonized parcel data, and initiatives such as the Ohio Geographically Referenced Information Program (OGRIP) and MassGIS seek to make statewide parcel data publicly available.[38,39] Adding features tailored to specific health conditions could refine the model's utility for targeted interventions. For instance, incorporating greenspace or air pollution measures could be used for respiratory-related hospitalizations, while proximity to airports or major highways can inform lead poisoning risk assessments. Other potential enhancements include identifying rental properties, including linkage to landlords, and integrating address-level data on evictions or property renovations, which can be indicators of housing instability and in-home environmental hazards. The model could be further specialized to assess risk at rental properties or in specific neighborhoods. Additionally, children in foster care, a high-risk population, were likely underrepresented in our dataset due to address filtering that excluded hospitalization records associated with Hamilton County Job and Family Services (HCJFS), limiting our ability to fully assess their risk. Future work could include partnership with HCJFS to more accurately link admissions to foster children's addresses of residence. Further, this modeling framework could be applied to ED visits, preventive care completion, or total cost of care as potential outcomes alongside hospitalizations. While this would require different

interpretation, it could provide valuable insights into broader patterns of healthcare utilization.

Further model development, calibration, and communication should incorporate feedback from partners and end-users to ensure that outputs are actionable, interpretable, and trustworthy. Operationalization of outputs and supporting interpretability will be critical to integrating this tool into real-world settings. Finally, validation at the patient level should be assessed with a prospective healthcare cohort to evaluate overall performance and model fairness.

By integrating open address-level data with hospital utilization records in a machine learning model, we have demonstrated the potential of a novel framework for identifying health risks with high spatial precision without compromising patient privacy.

ACKNOWLEDGEMENTS

Births per address data used in this study were obtained from the Bureau of Vital Statistics, Ohio Department of Health (ODH). Use of these data does not imply ODH agrees or disagrees with any presentations, analyses, interpretations, or conclusions.

COMPETING INTERESTS

The authors declare none.

FUNDING

This work was supported by grant funding from the Agency for Healthcare Research and Quality (R01HS027996) and the National Library of Medicine (R01LM013222).

FIGURES

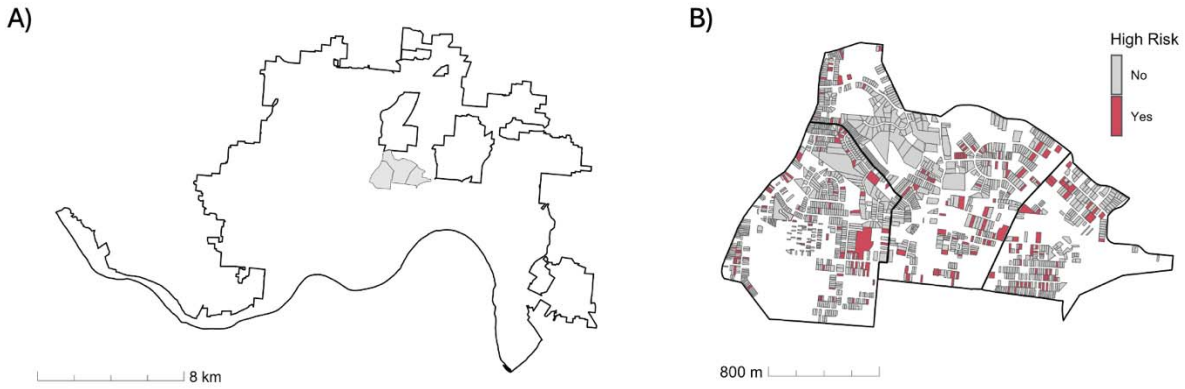

Figure 1. A map of census tract and residential property geographies in Cincinnati, Ohio. A) Three census tracts located directly northeast of Cincinnati Children's within the city of Cincinnati. According to 2019 American Community Survey 5-year averages, these tracts are comprised of a total population of 9,656 people in 4,173 households, including 3,055 children across 1,148 households. B) Geographies of 1,582 residential properties representing 1,799 addresses within the same three census tracts. Properties associated with high-risk addresses as classified by the hospitalization risk model at the top 7.4% diagnostic threshold are colored red.

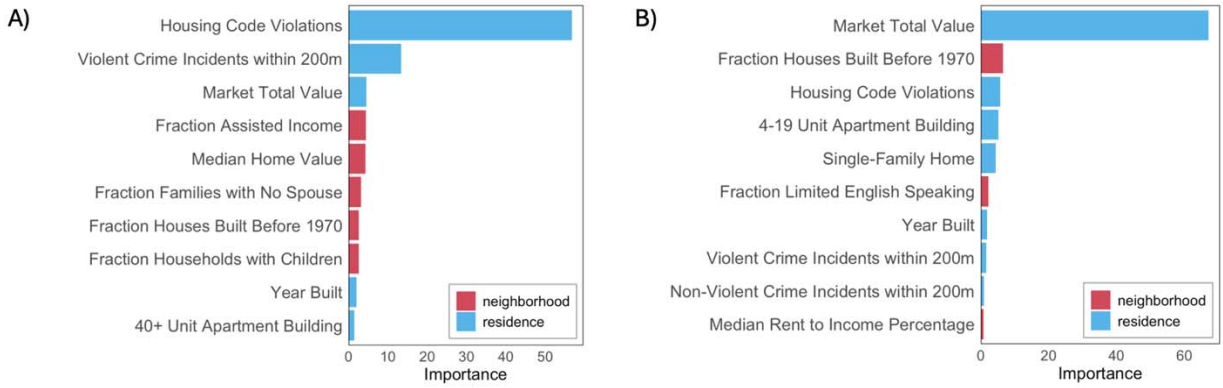

Figure 2. Variable importance plots. A) Top 10 most important features in the

hospitalization risk model. B) Top 10 most important features in the birth-adjusted hospitalization risk model.

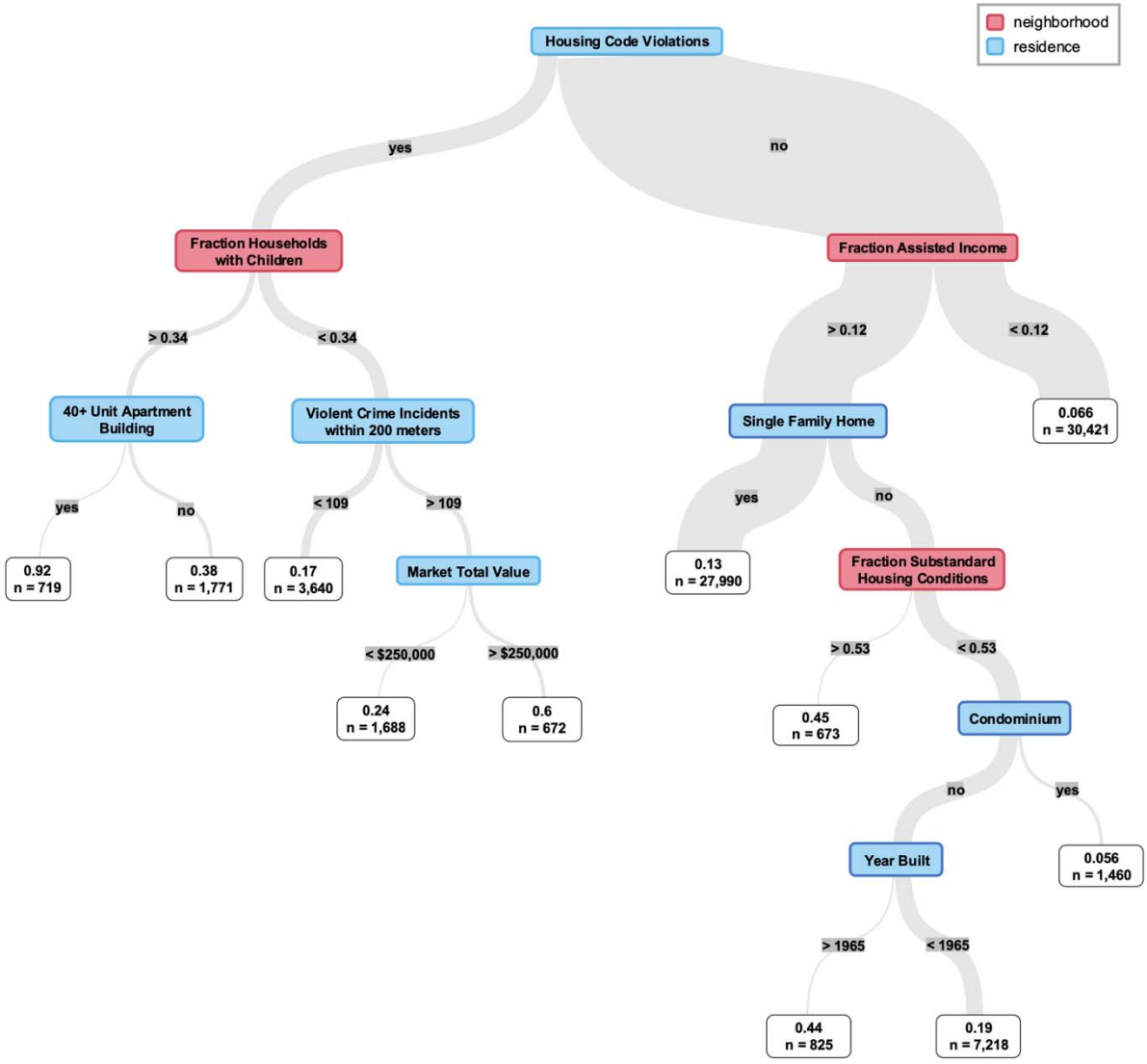

Figure 3. Decision tree fit to the output scores of the hospitalization risk model.

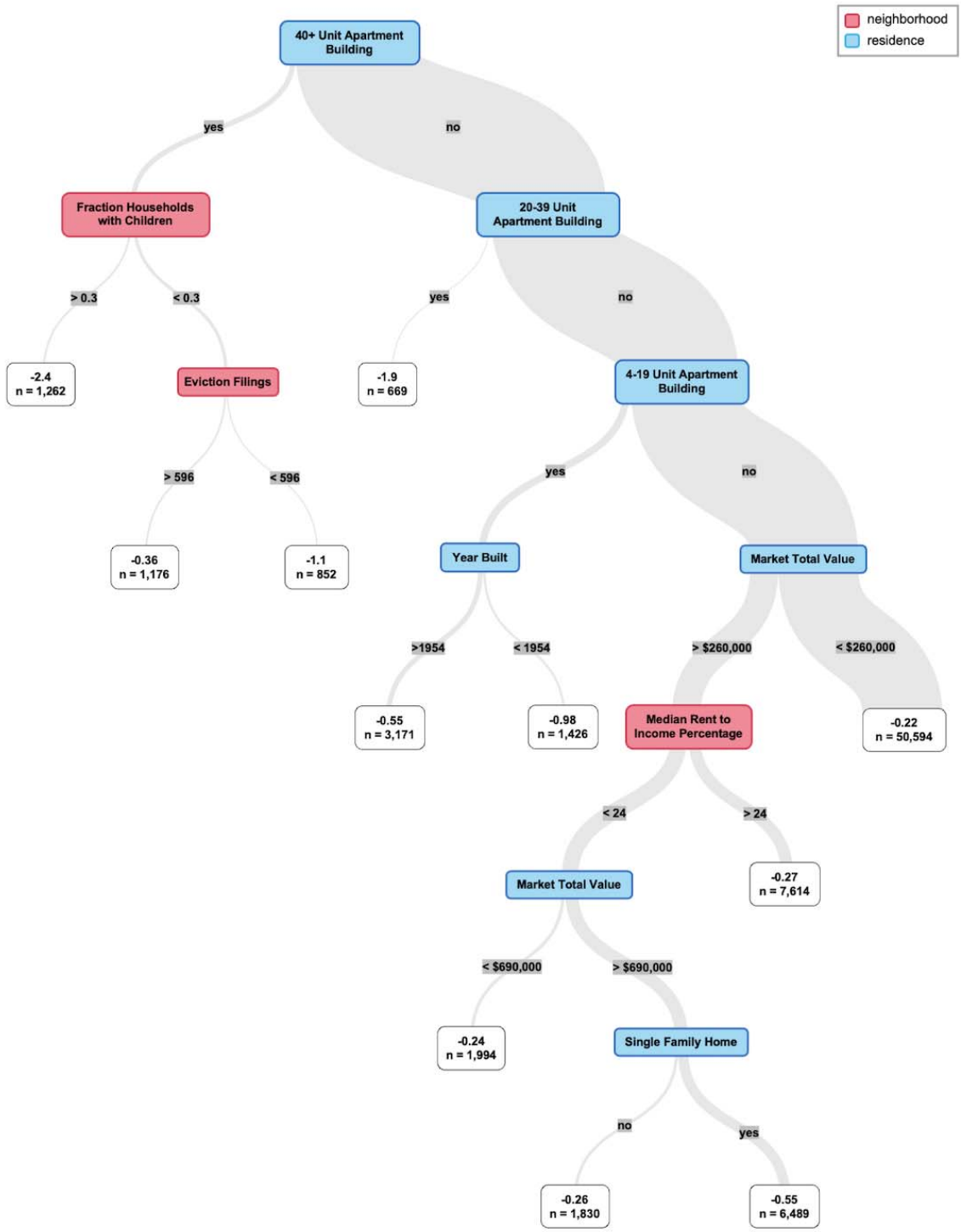

Figure 4. Decision tree fit to the output scores of the birth-adjusted hospitalization risk model.

REFERENCES

1. Beck AF, Riley CL, Taylor SC, Brokamp C, Kahn RS. Pervasive income-based disparities in inpatient bed-day rates across conditions and subspecialties. *Health Aff (Millwood)*. 2018;37(4):551-559. doi:10.1377/hlthaff.2017.1280

2. Nelson CA, Bhutta ZA, Harris NB, Danese A, Samara M. Adversity in childhood is linked to mental and physical health throughout life. *BMJ*. 2020;371:m3048. doi:10.1136/bmj.m3048

3. Cammarata C, Bujoreanu S, Wohlheiter K. Hospitalization and Its Impact: Stressors Associated with Inpatient Hospitalization for the Child and Family. In: Carter BD, Kullgren KA, eds. *Clinical Handbook of Psychological Consultation in Pediatric Medical Settings*. Springer International Publishing; 2020:37-49. doi:10.1007/978-3-030-35598-2_4

4. Bassett HK, Coller RJ, Beck J, et al. Financial Difficulties in Families of Hospitalized Children. *Journal of Hospital Medicine*. 2020;15(11):652-658. doi:10.12788/jhm.3500

5. Pardo-Crespo MR, Narla NP, Williams AR, et al. Comparison of individual-level versus area-level socioeconomic measures in assessing health outcomes of children in Olmsted County, Minnesota. *J Epidemiol Community Health*. 2013;67(4):305-310. doi:10.1136/jech-2012-201742

6. Manning ER, Duan Q, Taylor S, et al. Development of a multimodal geomarker pipeline to assess the impact of social, economic, and environmental factors on pediatric health outcomes. *Journal of the American Medical Informatics Association*. Published online May 10, 2024:ocae093. doi:10.1093/jamia/ocae093

7. Laurent O, Hu J, Li L, et al. A Statewide Nested Case–Control Study of Preterm Birth and Air Pollution by Source and Composition: California, 2001–2008. *Environ Health Perspect*. 2016;124(9):1479-1486. doi:10.1289/ehp.1510133

8. Hartlage CS, Manning ER, Brokamp C. Evaluating linkage approaches for address-level socioenvironmental exposure assessment. *medRxiv*. Preprint posted online December 4, 2025:2025.12.02.25341378. doi:10.64898/2025.12.02.25341378

9. Suárez-Paniagua V, Casey A, Marwick CA, et al. Care home resident identification: A comparison of address matching methods with Natural Language Processing. *PLOS ONE*. 2024;19(12):e0309341. doi:10.1371/journal.pone.0309341

10. Wilson JS, Shepherd DC, Rosenman MB, Kho AN. Identifying risk factors for healthcare-associated infections from electronic medical record home address data. *International Journal of Health Geographics*. 2010;9(1):47. doi:10.1186/1476-072X-9-47

11. Cutuli JJ, McDuffie MJ, Nescott E. Evaluating Approaches to Linking Evictions Records: Assessing the Feasibility of Research with Integrated Data. *Dela J Public Health*. 2023;9(2):24-29. doi:10.32481/djph.2023.06.006

12. Child and Adolescent Health Measurement Initiative. 2022-2023 National Survey of Children's Health (NSCH) data query. Data Resource Center for Child and Adolescent Health. Accessed May 20, 2025. https://mchb.hrsa.gov/data-research/national-survey-childrens-health

13. Foryciarz A, Gladish N, Rehkopf DH, Rose S. Incorporating area-level social drivers of health in predictive algorithms using electronic health record data. *Journal of the American Medical Informatics Association*. 2025;32(3):595-601. doi:10.1093/jamia/ocaf009

14. Cincinnati Area Geographic Information System (CAGIS). Accessed January 3, 2025. https://cagismaps.hamilton-co.org/cagisportal

15. Manning E, Brokamp C. addr. Published online September 23, 2025. Accessed January 3, 2025. https://github.com/geomarker-io/addr

16. Cincinnati Statistical Neighborhood Approximations (SNA) 2020. City of Cincinnati. July 26, 2023. Accessed January 3, 2025. https://data.cincinnati-oh.gov/dataset/Cincinnati-Statistical-Neighborhood-Approximations/i9zh-juvu/about_data

17. Manning E, Brokamp C. parcel. Published online 2024. Accessed May 29, 2025. https://github.com/geomarker-io/parcel

18. Hamilton County Auditor. Hamilton County Auditor. Accessed January 3, 2025. https://www.hamiltoncountyauditor.org/

19. PDI (Police Data Initiative) Crime Incidents | Tyler Data & Insights. City of Cincinnati. January 3, 2025. Accessed January 3, 2025. https://data.cincinnati-oh.gov/safety/PDI-Police-Data-Initiative-Crime-Incidents/k59e-2pvf/about_data

20. Cincinnati, OH. Eviction Lab. December 1, 2024. Accessed January 3, 2025. https://evictionlab.org/eviction-tracking/cincinnati-oh/

21. Athey S, Tibshirani J, Wager S. Generalized random forests. *The Annals of Statistics*. 2019;47(2):1148-1178. doi:10.1214/18-AOS1709

22. Rehm CD, Moudon AV, Hurvitz PM, Drewnowski A. Residential property values are associated with obesity among women in King County, WA, USA. *Soc Sci Med*. 2012;75(3):491-495. doi:10.1016/j.socscimed.2012.03.041

23. Drewnowski A, Buszkiewicz J, Aggarwal A, Cook A, Moudon AV. A new method to visualize obesity prevalence in Seattle-King County at the census block level. *Obesity Science & Practice*. 2018;4(1):14-19. doi:10.1002/osp4.144

24. Hughes AE, Tiro JA, Balasubramanian BA, Skinner CS, Pruitt SL. Social Disadvantage, Healthcare Utilization, and Colorectal Cancer Screening: Leveraging Longitudinal Patient Address and Health Records Data. *Cancer Epidemiol Biomarkers Prev*. 2018;27(12):1424-1432. doi:10.1158/1055-9965.EPI-18-0446

25. Liévanos RS, Evans CR, Light R. An Intercategorical Ecology of Lead Exposure: Complex Environmental Health Vulnerabilities in the Flint Water Crisis. *Int J Environ Res Public Health*. 2021;18(5):2217. doi:10.3390/ijerph18052217

26. Samuels EA, Taylor RA, Pendyal A, et al. Mapping emergency department asthma visits to identify poor-quality housing in New Haven, CT, USA: a retrospective cohort study. *Lancet Public Health*. 2022;7(8):e694-e704. doi:10.1016/S2468-2667(22)00143-8

27. Robb K, Ahmed R, Wong J, Ladd E, de Jong J. Substandard housing and the risk of COVID-19 infection and disease severity: A retrospective cohort study. *SSM - Population Health*. 2024;25:101629. doi:10.1016/j.ssmph.2024.101629

28. Chakraborty O, Dragan KL, Ellen IG, et al. Housing-Sensitive Health Conditions Can Predict Poor-Quality Housing. *Health Affairs*. 2024;43(2):297-304. doi:10.1377/hlthaff.2023.01008

29. Bozigar M, Connolly CL, Legler A, et al. In-home environmental exposures predicted from geospatial characteristics of the built environment and electronic health records of children with asthma. *Annals of Epidemiology*. 2022;73:38-47. doi:10.1016/j.annepidem.2022.06.034

30. Santos FO dos, Conti S, Wolters A. A Novel Method for Identifying Care Home Residents in England: A Validation Study. *International Journal of Population Data Science*. 2020;5(4). doi:10.23889/ijpds.v5i4.1666

31. Meeker JR, Burris H, Boland MR. An algorithm to identify residential mobility from electronic health-record data. *Int J Epidemiol*. 2022;50(6):2048-2057. doi:10.1093/ije/dyab064

32. Manning ER, Duan Q, Brokamp C. Incorporating Parcel-Based Housing Conditions to Increase the Precision of Identifying Children With Elevated Blood Lead. *J Public Health Manag Pract*. Published online January 17, 2025. doi:10.1097/PHH.0000000000002109

33. Beck AF, Henize AW, Qiu T, et al. Reductions In Hospitalizations Among Children Referred To A Primary Care–Based Medical-Legal Partnership. *Health Affairs*. 2022;41(3):341-349. doi:10.1377/hlthaff.2021.00905

34. Feldman M, Friedler S, Moeller J, Scheidegger C, Venkatasubramanian S. Certifying and Removing Disparate Impact. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Published online January 1, 2015. https://scholarship.haverford.edu/compsci_facpubs/116

35. Makkad H, Saini A, Manning ER, Duan Q, Colegate S, Brokamp C. Racial Fairness of Individual- and Community-Level Proxies of Socioeconomic Status Among Birthing Parent-Child Dyads. *J Racial Ethn Health Disparities*. Published online June 25, 2024. doi:10.1007/s40615-024-02050-9

36. Kontokosta CE, Hong B. Bias in smart city governance: How socio-spatial disparities in 311 complaint behavior impact the fairness of data-driven decisions. *Sustainable Cities and Society*. 2021;64:102503. doi:10.1016/j.scs.2020.102503

37. Lemire E, Samuels EA, Wang W, Haber A. Unequal Housing Conditions And Code Enforcement Contribute To Asthma Disparities In Boston, Massachusetts. *Health Affairs*. 2022;41(4):563-572. doi:10.1377/hlthaff.2021.01403

38. Ohio Parcels. Ohio Geographically Referenced Information Program. Accessed December 18, 2025. https://ohioparcels-geohio.hub.arcgis.com/

39. Commonwealth of Massachusetts. MassGIS (Bureau of Geographic Information). Mass.gov. 2025. Accessed November 24, 2025. https://www.mass.gov/orgs/massgis-bureau-of-geographic-information