



Harnessing the Power of a Novel Synthetic U.S. Population to Uncover Public Health Insights

Citation

Parikh, Shivani. 2025. Harnessing the Power of a Novel Synthetic U.S. Population to Uncover Public Health Insights. Doctoral Dissertation, Harvard University Graduate School of Arts and Sciences.

Link

<https://dash.harvard.edu/handle/1/42719400>

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

<https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFIOWI2M2EwYTg>

Accessibility

<https://accessibility.huit.harvard.edu/digital-accessibility-policy>

Share Your Story

The Harvard community has made this article openly available.

Please share how this access benefits you. [Submit a story](#)

HARVARD

Kenneth C. Griffin



GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Department of Population Health Sciences,
have examined a dissertation, entitled,

“Harnessing the Power of a Novel Synthetic U.S. Population to Uncover Public Health Insights”

presented by

SHIVANI PARIKH

candidate for the degree of Doctor of Philosophy
and hereby certify that it is worthy of acceptance.

Joseph Allen

Joseph Allen (Jan 6, 2025 13:28 EST)

Dr. Joseph G. Allen, Sc.D., Committee Chair, Harvard T.H. Chan School of Public Health

Brent A Coull

Brent A Coull (Jan 6, 2025 13:31 EST)

Dr. Brent A. Coull, Ph.D., Harvard T.H. Chan School of Public Health

John D Spengler

John D Spengler (Jan 8, 2025 06:51 PST)

Dr. John D. Spengler, Ph.D., Harvard T. H. Chan School of Public Health, Harvard Faculty of Arts and Sciences

Date: 19 December 2024

**Harnessing the Power of a Novel Synthetic U.S. Population
to Uncover Public Health Insights**

A dissertation presented
by
Shivani Parikh
to
the Population Health Sciences Program
in the Environmental Health Field of Study

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Harvard University
Cambridge, Massachusetts

December 2024

© 2024 Shivani Parikh

All rights reserved

ABSTRACT

Marginalized communities in the U.S. must contend with structural discrimination that has systematically placed them in under-resourced neighborhoods whose inherent social, economic, infrastructural, and environmental conditions adversely affect health and their wellbeing. Today, researchers' understanding of these neighborhoods—characterized by substandard housing, hazardous workplaces, degraded infrastructures, close proximity to toxic sites and polluting facilities, food deserts and fast-food jungles, limited opportunities for quality education and employment, and restricted access to healthcare services and facilities—and their inhabitants rely on incomplete population data fraught with assumptions and biases symptomatic of a long history of residential segregation. Broadly, the two options for available population data are: (1) geographically-aggregated data that operates at a high spatial resolution, but only provides marginal distributions of single characteristics, or (2) individual-level microdata that provides joint distributions of multiple characteristics, but only operates at a low spatial resolution. Thus, researchers aiming to leverage available population data to devise and disperse public health interventions must forgo either an understanding of compounding risks or the ability to pinpoint localized risks.

This dissertation aims to construct, validate, and demonstrate the utility of a novel synthetic U.S. population which models 330 million agents that collectively reflect the real population with respect to key social, economic, infrastructural, and environmental characteristics at the census tract level. Generated through spatial microsimulation methods that avail the strengths of both geographically-aggregated data and individual-level microdata, it affords researchers the ability to answer previously unanswerable questions about large-scale, complex public health issues across the U.S. First, we constructed the model using iterative proportional fitting (IPF)—a widely-favored static deterministic reweighting method for spatial microsimulation. We then validated it against the

2018-2022 American Community Survey (ACS), the 2020 Decennial Census, and the 2020 Residential Energy Consumption Survey (RECS) with a suite of metrics including Pearson's R^2 (mean = 0.98), standard absolute error (SAE) (mean = 0.03), and the Kullback-Leibler (KL) divergence statistic (mean < 0.08), finding that our synthetic U.S. population exhibited strong overall performance in simulating the real U.S. population with variation due to the size and diversity of census tracts and the complexity of linking variables. Second, we examined the well-established compounding risk faced by children under the age of five who we know—based on our synthetic U.S. population—live in homes built before 1980 to map residential lead exposure hotspots at the census tract level. Our novel approach for identifying at-risk areas therefore employed the joint distribution of these two characteristics in direct contrast to traditional approaches that assess risk by overlapping separate marginal distributions. This enabled us to uncover up to 73% more at-risk census tracts where an estimated 3.28 million additional at-risk children reside. Additionally, our approach revealed statistically significant systematic inequities in how traditional approaches detect risk, where newly identified at-risk children were disproportionately located in census tracts with higher concentrations of uninsured, White, and Native populations. It also highlighted that relying on housing age as a proxy for residential lead-based paint exposure may neglect important disparities in housing quality. Third, we coupled our synthetic U.S. population with the 2020 Residential Energy Consumption Survey (RECS) to estimate individual probabilities of residential air conditioning (AC) access—an essential cooling solution to protect against the adverse effects of heat stress—using a machine learning model to downscale data that was previously only available with spatial information at the state level to the census tract level. Unlike traditional approaches which focus solely on AC ownership as a proxy for AC access, our measure additionally accounted for the ability to install, maintain, and operate AC, resulting in probability estimates that were 17% lower and 9% more variable. By accounting for these additional dimensions of residential AC access, we

also uncovered interstate and intrastate disparities shaped by fine-scale differences in state policies, infrastructures, and practices as well as climate conditions, housing conditions, and demographic characteristics.

In summary, our novel synthetic U.S. population addresses critical gaps in existing population data, uncovering limitations, assumptions, and biases that may otherwise persist as researchers aim to devise and disperse public health interventions. Beyond its utility in mapping residential lead exposure risk and residential AC access, this model can be adapted to examine a wide range of public health challenges, enabling researchers to explore how intersecting social, economic, infrastructural, and environmental forces shape outcomes at fine spatial resolutions and providing them with a powerful tool to guide more effective and equitable public health interventions across the U.S.

TABLE OF CONTENTS

TITLE PAGE.....	I
COPYRIGHT.....	II
ABSTRACT	III
TABLE OF CONTENTS.....	VI
ACKNOWLEDGEMENTS.....	IX
DEDICATION.....	XII
LIST OF TABLES.....	XIII
LIST OF FIGURES	XIV
LIST OF ACRONYMS AND ABBREVIATIONS.....	XVI
CHAPTER 1: PREFACE.....	1
CHAPTER 2: CONSTRUCTING A SYNTHETIC POPULATION OF THE UNITED STATES TO ENABLE PUBLIC HEALTH RESEARCH	9
1. ABSTRACT.....	10
2. INTRODUCTION.....	11
3. METHODS.....	18
3.1. Spatial microsimulation.....	19
3.1.1. Modeling framework.....	19
3.1.2. Linking variables.....	22
3.1.3. Geographic scale.....	23
3.2. Data.....	25
3.2.1. Geographically-aggregated data.....	25
3.2.1.1. Data collection.....	25
3.2.1.2. Data processing.....	26
3.2.2. Individual-level microdata.....	28
3.2.2.1. Data collection.....	28
3.2.2.2. Data processing.....	29
3.3. Analysis.....	33
3.4. Validation study.....	35
3.4.1. Internal validation.....	35
3.4.2. External validation.....	37
4. RESULTS.....	38
4.1. Synthetic U.S. population characteristics.....	39
4.2. Validation study.....	43
4.2.1. Visual Validation.....	44
4.2.2. Internal validation.....	53
4.2.3. External validation.....	55
5. DISCUSSION.....	58

5.1. Strengths.....	61
5.2. Limitations.	62
6. CONCLUSIONS.....	64
CHAPTER 3: LEVERAGING A SYNTHETIC POPULATION TO IDENTIFY CHILDREN AT HIGH RISK OF RESIDENTIAL LEAD EXPOSURE ACROSS THE UNITED STATES..66	
1. ABSTRACT.....	67
2. INTRODUCTION.....	68
3. METHODS.....	74
3.1. Data sources.....	74
3.2. Statistical analysis.	75
3.2.1. Comparing census-tract-level risk classification.	75
3.2.1.1. Quantifying the extent of exposure misclassification.	76
3.2.1.2. Quantifying the nature of exposure misclassification.....	77
3.2.2. Identifying individual-level risks of young children living in older homes.	79
4. RESULTS.....	81
4.1. Comparing census-tract-level risk classification.....	81
4.1.1. Quantifying the extent of exposure misclassification.	83
4.1.2. Quantifying the nature of exposure misclassification.....	84
4.2. Identifying individual-level risks of young children living in older homes.	85
5. DISCUSSION.....	89
5.1. Strengths.....	98
5.2. Limitations.	99
6. CONCLUSIONS.....	101
CHAPTER 4: LEVERAGING A SYNTHETIC POPULATION TO IDENTIFY POPULATIONS UNPROTECTED AGAINST HEAT STRESS ACROSS THE UNITED STATES.....103	
1. ABSTRACT.....	104
2. INTRODUCTION.....	105
3. METHODS.....	108
3.1. Data sources.....	110
3.1.1. RECS dataset.	110
3.1.2. Synthetic U.S. population dataset.	111
3.2. Data processing.....	112
3.2.1. Selecting the target variable for AC access from the RECS dataset.....	112
3.2.2. Standardizing the predictive variables across the RECS and synthetic U.S. population datasets.	114
3.2.3. Harmonizing the underlying structures of the RECS and synthetic U.S. population datasets.	115
3.3.3. Balancing the two AC access classes.	118
3.3. Modeling.....	119
3.3.1. Model training and validation.....	120
3.3.2. Model application.....	121
3.4. Analysis.....	122
4. RESULTS.....	122
4.1 Study population.	122
4.2 Key predictors of residential AC access.	125

4.3. Assessing the spatial distribution of residential AC access.....	126
4.4. Assessing the influence of tract-level characteristics on residential AC access.....	128
4.5. Assessing the influence of individual-level characteristics on residential AC access.	129
5. DISCUSSION.....	133
5.1. Strengths.....	143
5.2. Limitations.	144
6. CONCLUSIONS.....	147
CHAPTER 5: CONCLUSIONS.....	149
Summary of findings.....	149
Practical implications of findings.....	150
Recommendations for future research.....	153
Concluding remarks	156
APPENDIX.....	158
Appendix A.	158
Appendix B.....	161
REFERENCES.	164

ACKNOWLEDGEMENTS

It takes a village, and I want to begin by expressing my gratitude to the extraordinary group of mentors, colleagues, friends, family members, and loved ones who have supported me throughout this journey.

First and foremost, I owe an immeasurable debt of gratitude to my advisor and the chair of my Dissertation Advisory Committee, Dr. Joseph Allen. Over the past five years, his mentorship has shaped the trajectory of my life, opening doors I never expected and inspiring me to pursue paths I never imagined to be possible. Dr. Allen's clear passion for healthy buildings and his unwavering commitment to leveraging them as a tool to improve lives on a global scale are nothing short of infectious. Central to his work is a guiding ethos captured in the question that has been displayed on the wall of his lab for the past ten years: *"How will your research impact the world?"* This ethos not only defines his approach to his research but also sets the standard he holds for his team. It has transformed the way I approach my own work, instilling in me the belief that every decision made throughout the research process—down to the smallest technical detail—should always serve the larger goal of creating meaningful, lasting change in the world.

Dr. Brent Coull and Dr. John Spengler, who also serve as members of my Dissertation Advisory Committee, have played a significant role in my professional growth. Their probing questions and insightful suggestions have not only shaped my research but also taught me that the true value of a PhD lies in cultivating critical thinking and problem-solving skills. I am especially grateful to Dr. Coull for sharing his technical expertise, which he approaches with remarkable thoughtfulness and care, making even the most complex concepts accessible. His kindness, generosity, and enthusiasm for statistical methods have shown me how powerful—and even fun—they can be in solving real-world problems. I am equally grateful to Dr. Spengler, whose work has been instrumental in advancing the healthy buildings movement. His incisive questions about how

to translate my research into tangible policies, programs, and practices have pushed me to ensure that it is rooted in real-world applicability and poised to make the greatest impact. That same focus is reflected in his commitment to fostering collaboration across people, disciplines, and sectors, reinforcing the idea that collective effort and shared vision are essential for addressing complex public health challenges.

In addition to their guidance, I want to acknowledge Dr. David Christiani for serving on my Dissertation Proposal Qualifying Examination Committee alongside Dr. Coull and Dr. Spengler, as well as for facilitating funding support through the National Institute for Occupational Safety and Health grant T42 OH008416, which sustained my work throughout the doctoral program.

I also want to thank Dr. Nancy Krieger for her invaluable guidance. She has taught me to critically examine the assumptions and biases of the people who conduct research, including my own, and to see science as deeply interconnected with the world rather than separate from it. Her example has also fostered my love for teaching and the reciprocal learning process it creates with students. Dr. Krieger's influence has left a lasting impact on how I approach both research and pedagogy, and I feel fortunate to have learned from her.

I am grateful to each and every member of the Healthy Buildings team—past and present—who have supported and inspired me throughout my time at Harvard. In particular, I want to thank Dr. Anna Young and Dr. Lauren Ferguson, who have been endlessly encouraging and who I will forever consider as among the best mentors I have had the privilege to learn from. They have guided me through every stage of the research process, from generating questions and designing studies to analyzing data and communicating findings, and I am proud to now call them dear friends. I am also grateful to Dr. Allen's past doctoral students—Dr. Piers MacNaughton, Dr. Anna Young, Dr. Erika Eitland, Dr. Emily Jones, and Dr. Sandra Dedesko—whose sage wisdom and sound advice have guided me through this journey. Finally, I want to thank Maya Sands-Bliss, Mahala

Lahvis, Rachel Steiner, Amanda Hernandez, Kate Harson, Sophie Erlich, Seetha Davis, and Rienna Russo, who have not only supported my work but have also provided me with community in the workplace when I needed it most.

To close, I want to express my heartfelt appreciation for my friends and family, who have brought so much meaning to my life. To my parents, Bhairavi and Rajiv Parikh, who have both kept me grounded and instilled in me the values of curiosity, hard work, and courage to pursue my passions. To my siblings, Arjun, Anjali, and Anand Parikh, who have broadened my perspective through their own beliefs, convictions, and passions across law, engineering, and computer science. To my life partner, Noah Jordan Parikh Cott, whose unwavering support each step of the way has meant more to me than I can put into words. He has taught me the importance of balance—of stepping back, spending time with loved ones, and embracing the joys of life outside of work—practices that have not only enriched my life but also strengthened my work. For his love, encouragement, patience, and partnership, I am endlessly grateful. And to my little rascal puppy, Sully, who has never failed to bring a smile to my face. Thank you.

“My mother never gave me any idea that I couldn’t do whatever I wanted to do or be whomever I wanted to be.

I don’t know if she ever realized that the person I most wanted to be was her.” (Gilmore Girls, 2003)

LIST OF TABLES

Table 1. Geographically-aggregated data – “wide” structure.....	26
Table 2. Individual-level microdata – “long” structure.....	29
Table 3. Comparison of variable sub-categories across input datasets.....	31
Table 4. Goodness-of-fit metrics for internal validation.....	36
Table 5. Synthetic population dataset – “long” structure.....	38
Table 6. Descriptive statistics of the synthetic U.S. population dataset.....	39
Table 7. Internal validation results.....	54
Table 8. Results of the SAR model presented for the full population.....	85
Table 9. Descriptive statistics of children under the age of six in the synthetic U.S. population dataset.....	86
Table 10. Results of the GLMM presented for the full population and stratified by poverty level.....	88
Table 11. Aligning the structures of the RECS and the synthetic U.S. population datasets to facilitate predictive modeling.....	116
Table 12. Descriptive statistics of the synthetic U.S. population dataset.....	123

LIST OF FIGURES

Figure 1. Dissertation framework.....	5
Figure 2. Schematic diagram of spatial microsimulation.....	16
Figure 3. Methodological framework for spatial microsimulation.....	21
Figure 4. Comparison of linking variable distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in the most populous county in each state and the least populous county in each state.....	46
Figure 5. Pearson’s R versus population size by census tract.....	55
Figure 6. KL divergence statistic results for age, sex, race, and ethnicity at the census tract level, and housing tenure, housing occupancy, building type, building age, and heating fuel at the state level.....	57
Figure 7. Visual comparison of census tracts identified as “high-risk” for lead exposure across the U.S. by the traditional marginal distribution approach using the ACS dataset versus our novel joint distribution approach using the synthetic U.S. population dataset.....	82
Figure 8. Percent of tracts not detected as “high-risk” for childhood residential lead exposure by the traditional marginal distribution approach compared to our novel joint distribution ordered by state.....	84
Figure 9. Concentration of uninsured children under the age of six living in homes built before 1980 at the census tract level across the U.S.....	94
Figure 10. Concentration of Native American children under the age of six living in homes built before 1980 at the census tract level across the U.S.....	96
Figure 11. Concentration of impoverished Black children under the age of six living in homes built before 1980 at the census tract level across the U.S.....	97
Figure 12. Concentration of impoverished White children under the age of six living in homes built before 1980 at the census tract level across the U.S.....	98
Figure 13. Methodological framework.....	110
Figure 14. Predicted probability of limited residential AC access modeled as a joint function of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDDs, and state using microdata drawn from RECS at the state level versus microdata drawn from the synthetic U.S. population at the census tract level.....	127
Figure 15. Predicted probability of limited residential AC access by census-tract-level characteristics, including annual CDD population density, and the percentage of the population that is both non-White and below the federal poverty level.....	129

Figure 16. Predicted probability of limited residential AC access by individual-level characteristics, including race ethnicity, household income, housing tenure, building type, building age, housing occupancy.....132

Figure 17. Predicted probability of limited residential AC access in WA using microdata drawn from RECS at the state level versus microdata drawn from the synthetic U.S. population at the census tract level.....141

Figure 18. Predicted probability of limited residential AC access in NY using microdata drawn from RECS at the state level versus microdata drawn from the synthetic U.S. population at the census tract level.....142

Figure 19. Predicted probability of limited residential AC access in NY using microdata drawn from RECS at the state level versus microdata drawn from the synthetic U.S. population at the census tract level.....143

LIST OF ACRONYMS AND ABBREVIATIONS

AC: Air Conditioning

ACS: American Community Survey

AHS: American Housing Survey

APT: Apartment

ATSDR: Agency for Toxic Substances and Disease Registry

AUC: Area Under the Curve

BLL: Blood Lead Level

CBSA: Core-Based Statistical Area

CDC: Centers for Disease Control and Prevention

CDD: Cooling Degree Days

CDO: Climate Data Online

CEJST: White House Climate and Economic Justice Screening Tool

CT: Connecticut

DOE: Department of Energy

EIA: Energy Information Administration

EJI: Environmental Justice Index

EPA: Environmental Protection Agency

ESS: Energy Supplier Survey

FBS: Fitness-Based Synthesis

FDA: Food and Drug Administration

FEMA: Federal Emergency Management Agency

FIPS: Federal Information Processing Standards

FPL: Federal Poverty Level

GLMM: Generalized Linear Mixed Model

HDD: Heating Degree Days

HEAP: Home Energy Assistance Program

HHS: Department of Health and Human Services

HUD: Department of Housing and Urban Development

IHA: Indian Housing Assistance

IHBG: Indian Housing Block Grant

IPF: Iterative Proportional Fitting

IPU: Iterative Proportional Updating

IQR: Inter-Quartile Range

IRS: Internal Revenue Service

KL: Kullback-Leibler

LBP: Lead-Based Paint

LEAD: Low-Income Energy Affordability Data

LERI: Lead Exposure Risk Index

LIHEAP: Low Income Home Energy Assistance Program

NAHASDA: Native American Housing Assistance and Self-Determination Act

NCEI: National Centers for Environmental Information

NIMBY: Not in my back yard

NOAA: National Oceanic and Atmospheric Administration

NRDC: Natural Resources Defense Council

NREL: National Renewable Energy Laboratory

NRSME: Normalized Root Mean Squared Error

NY: New York

OECD: Organization for Economic Co-operation and Development

OR: Odds Ratio

OSHA: Occupational Safety and Health Administration

P-MEDM: Penalized Maximum Entropy

PTSD: Post-Traumatic Stress Disorder

PUMA: Public Use Microdata Area

PUMS: Public Use Microdata Sample

RAPT: Resilience Analysis and Planning Tool

RCP: Representative Concentration Pathway

RECS: Residential Energy Consumption Survey

RMSE: Root Mean Squared Error

ROC: Receiver Operating Characteristic

SAE: Standard Absolute Error

SAR: Spatial Autoregressive Model

SD: Standard Deviation

SFH: Single-Family Home

SLOPE: State and Local Planning for Energy

SMOTE: Synthetic Minority Oversampling Technique

SVI: Social Vulnerability Index

TAE: Total Absolute Error

TRS: Truncate, Replicate, Sample

TX: Texas

UK: United Kingdom

US: United States

USDA: United States Department of Agriculture

WA: Washington

WAP: Weatherization Assistance Program

CHAPTER 1: PREFACE

It is deeply troubling that the United States (U.S.)—a nation that prides itself on being a global leader in economic power and technological innovation—consistently ranks at the bottom among its peer nations in safeguarding the health and well-being of its people despite investing more into its healthcare system than any other nation in the world.¹ Americans face shockingly high rates of infant and maternal mortality, communicable and noncommunicable disease, mental illness, and injury fatality, resulting in a life expectancy that is three years below the Organization for Economic Co-operation and Development (OECD) average.¹

The core issue is that the U.S. healthcare system is fundamentally misaligned with the goal of optimizing population health. Driven by a profit-based model designed to prioritize high-cost, high-tech treatments over prevention, the system is plagued by fragmented care teams, the absence of universal healthcare coverage, and high out-of-pocket healthcare costs, forcing many to delay or forgo the care they need.¹⁻⁷ As a result, health conditions that could have been effectively managed with timely intervention often worsen unnecessarily, leading to the overuse of emergency room visits,⁸ hospitalizations,⁹ and complex medical procedures.¹⁰ Yet, while the U.S. healthcare system pours its resources into these reactionary treatments, it neglects the very basics of what actually keeps people healthy: clean air, safe water, nutritious foods, healthy buildings, safe neighborhoods, social support systems, in addition to accessible primary care.^{1,6-8,11-14} This misalignment drives up healthcare costs while failing to deliver the meaningful gains to population health that one would expect from such a high level of investment.

While these failures affect the entire population, research consistently shows that they weigh most heavily on marginalized communities—those historically subjected to structural (including interpersonal and institutional) discrimination on the basis of their race, ethnicity, immigration

status, Indigenous status, gender, sexuality, age, disability, and/or class.¹⁵ For example, as previously noted, the overall life expectancy in the U.S. is three years lower than the OECD average. However, this figure masks stark class-based disparities within the U.S., with life expectancy of those in the bottom 1% of income earners 10-15 years lower compared to those in the top 1%.¹⁶ Implicit bias in healthcare settings means Black women receive lower-quality healthcare, resulting in maternal mortality rates 3-5 times higher than those of White women,¹⁷⁻¹⁹ despite 80% of these deaths being preventable.²⁰ Language barriers, lack of healthcare coverage, and mistrust in the healthcare system—reinforced by fears of discrimination and/or deportation by U.S. institutions—are among the factors that deter Hispanic Americans from seeking vaccination and timely care, leading to disproportionately higher rates of illness and death from COVID-19 compared to non-Hispanic Americans.²¹⁻²³ Likewise, Native Americans who have endured intergenerational trauma from forced displacement, cultural suppression, and violence experience significantly higher rates of mental illness than the U.S. population.^{24,25} They are 2-3 times more likely to develop post-traumatic stress disorder (PTSD), 1.5-2.5 times more likely to suffer from alcohol dependence, and four times more likely to die by suicide, representing the highest suicide rate among any minoritized group in the U.S.²⁶⁻²⁸ These examples are just a few of many, illustrating a broader trend where structural discrimination not only produces health inequities but also perpetuates the failures of the U.S. healthcare system—a cyclical pattern seen across virtually every health outcome.

It would be short-sighted to view these failures solely as a matter of interpersonal and institutional discrimination within healthcare settings. The places where people live, work, learn, and play—in shaping exposures to risks, access to resources, and social, economic, and political support systems—are powerful determinants of health and health equity.^{29,30} Yet, the environments in which people find themselves are not by happenstance; they are products of historical and ongoing processes to systematically segregate specific subpopulations.³¹⁻³³ Historical policies such as

redlining, exclusionary zoning, restrictive covenants, and municipal underinvestment were explicitly designed to enforce racial, ethnic, and economic segregation.³³ Although these policies are no longer legal, their impacts persist today and are compounded by more implicit forms of systemic segregation, such as predatory lending and foreclosure practices, inequitable public infrastructure investment, local resistance to affordable housing (NIMBYism), and gentrification-driven displacement.³³ Together, these policies and practices have not only successfully segregated marginalized communities,³⁴ but have also concentrated them into under-resourced neighborhoods characterized by substandard housing, hazardous workplaces, degraded infrastructures, close proximity to toxic sites and polluting facilities, food deserts and fast-food jungles, limited opportunities for quality education and employment, and restricted access to healthcare services and facilities.^{29,30,35–40} A growing body of research has connected these deprived social and built environments with adverse health outcomes, including heightened vulnerability to experiencing acute symptoms, infectious disease, chronic conditions, mental illness, violence, injury, and premature death.^{29,30,39–45}

Therefore, in order to effectively address health and health inequities across the U.S., it is essential to understand both who is most vulnerable and where they live. However, researchers are constrained by the limitations inherent in the available population data; they must either rely on geographically-aggregated data that operates at a high spatial resolution, but only provides marginal distributions of single characteristics, or on individual-level microdata that provides joint distributions of multiple characteristics, but only operates at a low spatial resolution. In this way, researchers aiming to leverage available population data to devise and disperse public health interventions must forgo either an understanding of compounding risks or the ability to pinpoint localized risks. But as public health challenges grow increasingly complex—spanning widening inequities,^{46,47} infectious disease outbreaks,⁴⁸ climate change disasters,^{49–51} and ever-changing

migration patterns⁵²—the traditional usage of population data grows increasingly inadequate for analyzing the social, economic, infrastructural, and environmental interactions that form the basis of real-world population health dynamics. Instead, these so-called multifaceted, multilevel “wicked problems” demand the evolution of equally “wicked solutions.”⁵³

Spatial microsimulation is uniquely positioned to lead this evolution.⁵³ By integrating the joint distributions from individual-level microdata with the spatial resolution from geographically-aggregated data to generate a synthetic population, this approach permeates the barriers that have prevented us from accurately identifying the most vulnerable populations and meaningfully addressing their needs.⁵⁴ Synthetic populations not only support the simulation of interventions in a virtual environment, sidestepping the ethical and logistical challenges of large-scale real-world trials, but also provide a foundation for more advanced methodologies such as agent-based or future population simulations.⁵⁴ While these qualities make it an invaluable tool for planning, resource allocation, and policy development, most public health professionals have not yet tapped into its full potential.⁵⁵

To address this gap, this dissertation centers on harnessing the power of spatial microsimulation to uncover public health insights, following the framework outlined in **Figure 1**.

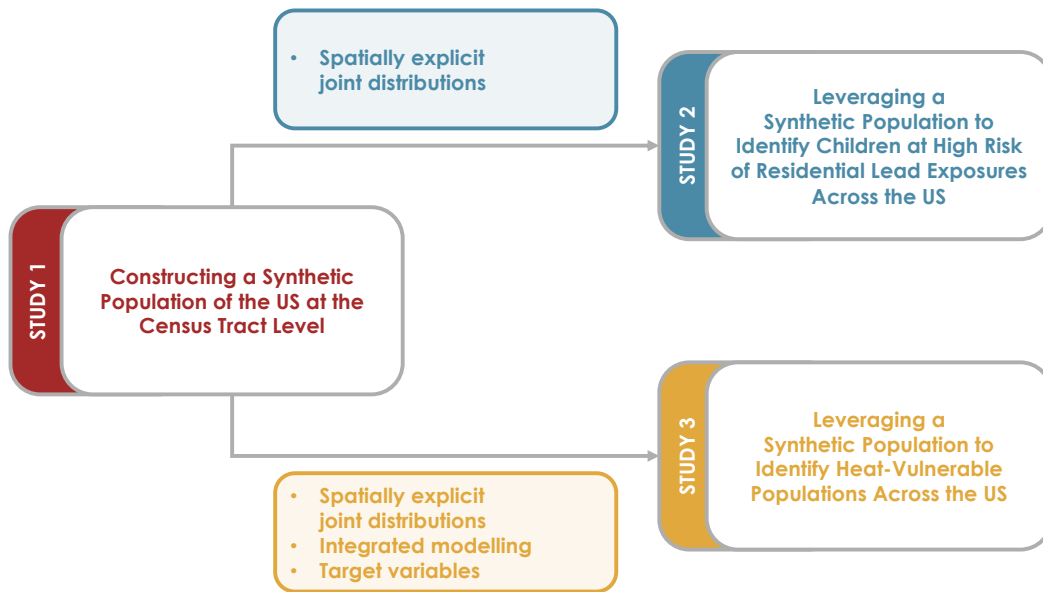


Figure 1. Dissertation framework.

In Study 1 (Chapter 2), we refine best practices and introduce key advancements in spatial microsimulation to construct an up-to-date synthetic U.S. population specifically tailored to public health research. We developed this synthetic U.S. population using a static deterministic reweighting method to constrain the joint distributions of linking variables from individual-level microdata—age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel type—to their known marginal distributions from geographically-aggregated data at the census tract level. Through extensive validation against “gold-standard” datasets, we demonstrate that our synthetic U.S. population reflects the real U.S. population with respect to the key social, economic, infrastructural, and environmental characteristics that drive population health dynamics at the local level. These features give our synthetic U.S. population three key capabilities that establish it as a foundational tool for unlocking new opportunities in public health research: (1) enabling the study of spatially explicit joint distributions; (2) facilitating the integration of external

datasets and models; and (3) leveraging these combined capabilities to gain access to target variables that are otherwise difficult to assess. In Studies 2 (Chapter 3) and 3 (Chapter 4), we demonstrate how these capabilities can be applied to answer two pressing public health questions, providing concrete examples of its utility in identifying at-risk populations and informing targeted interventions to mitigate long-standing disparities.

Study 2 employs the spatially explicit joint distributions in our synthetic U.S. population to address the century-long issue of young children facing compounding risks of lead exposure from living in older homes—a persistent and pervasive public health crisis with severe individual and societal costs. Despite regulatory bans on lead-based paint over 40 years ago, millions of homes built before 1978 across the U.S. still pose significant risks of lead exposure, particularly for children under the age of five whose developing bodies are especially vulnerable to lead’s toxic effects. A core priority of the Federal Lead Task Force’s action plan is to create residential lead exposure maps that identify high-risk areas in need of enhanced surveillance, care, and remediation. However, traditional mapping approaches to support these efforts identify high-risk census tracts by relying on overlapping marginal distributions of single characteristics, such as the percentage of children under the age of five or the percentage of homes built before 1980 in a given census tract. While useful, these approaches fail to capture how these characteristics co-occur, which can lead to the misidentification of high-risk areas and misallocation of resources. To address these limitations, Study 2 introduces a novel mapping approach enabled by our synthetic U.S. population to identify high-risk census tracts based on the co-occurrence of children under the age of five living in homes built before 1980. By comparing our novel joint distribution approach to the traditional marginal distribution approach, we revealed that up to 73% more at-risk census tracts containing up to 3.28 million additional at-risk children would have otherwise gone undetected. Statistical models further showed that traditional approaches systematically overlook at-risk children living in census tracts

with higher proportions of uninsured, White, and Native American populations and may overly rely on housing age as a proxy for residential lead exposure without accounting for disparities in housing quality. These findings demonstrate the unique capability of our synthetic U.S. population to uncover the nuanced interplay of social, economic, infrastructural, and environmental characteristics that drive localized vulnerabilities to residential lead exposure across the U.S., providing a more equitable framework for prioritizing lead mitigation efforts to the communities that need them most.

Study 3 leverages all three capabilities of our synthetic U.S. population to address the urgent need to understand disparities in residential air conditioning (AC) access—a critical strategy for protecting against the harmful effects of heat stress in the face of climate change. As global temperatures rise, limited AC access exacerbates vulnerability to heat-related illnesses and death, particularly among marginalized populations. However, the data required to identify who lacks AC access and where they live has been historically inadequate, hindering government programs including the Department of Energy’s (DOE) Weatherization Assistance Program (WAP) and the Department of Health and Human Services’ (HHS) Low Income Heat Energy Assistance Program (LIHEAP) from efficiently deploying cooling resources. Traditionally, researchers have relied on AC ownership data as a proxy for AC access, but this measure fails to capture other essential dimensions of AC access such as the ability to afford the high costs of AC installation, maintenance, and operation. While recently released data on AC access captures these dimensions, it only provides the state where each surveyed household is located, obscuring the intrastate variability that is critical for heat resilience planning and mitigation efforts typically implemented at the local level. To address these limitations, Study 3 introduces a novel integrated modeling approach that uses machine learning to combine microdata on AC access from the Residential Energy Consumption Survey (RECS) with microdata from our synthetic U.S. population, effectively downscaling data that was previously only available with spatial information at the state level to the census tract level. This

marks the first effort to estimate individual probabilities of AC access and evaluate how those probabilities vary by social, economic, infrastructural, and environmental characteristics at such a high spatial resolution across the entire U.S. By incorporating dimensions of AC access beyond ownership—such as the ability to afford installation, maintenance, and operational costs—we estimated probabilities that were on average 17% lower and 9% more variable than traditional estimates. Furthermore, our approach uncovered significant interstate and intrastate disparities driven by differences in state policies, infrastructure, and energy practices, as well as local climate conditions, housing conditions, and demographic characteristics—compounding vulnerabilities that cannot be captured by traditional approaches. Through this work, we demonstrate how our synthetic U.S. population can illuminate localized disparities in AC access across the U.S., offering the actionable insights necessary for researchers, planners, and policymakers to design effective cooling strategies that protect heat-vulnerable populations.

Overall, this dissertation underscores the transformative potential of our novel synthetic U.S. population to overcome the limitations, assumptions, and biases inherent in real population data traditionally employed in the field of public health, paving the way for a healthier, more sustainable, and more just future for the U.S. population.

CHAPTER 2: CONSTRUCTING A SYNTHETIC POPULATION OF THE UNITED STATES TO ENABLE PUBLIC HEALTH RESEARCH

Shivani Parikh^{a,b}, Lauren Ferguson^a, Brent A. Coull^a, John D. Spengler^a, Joseph G. Allen^a

^a Harvard T.H. Chan School of Public Health, 401 Park Drive, 4th Floor West, Boston, MA, 02215, USA

^b Harvard Graduate School of Arts and Sciences, 1350 Massachusetts Avenue, Cambridge, MA, 02138, USA

1. ABSTRACT.

Addressing public health challenges in the U.S.—where significant health disparities persist—requires high-resolution data that captures the complex interplay of social, economic, infrastructural, and environmental characteristics. However, strict data privacy regulations limit researchers to two main types of data: (1) individual-level microdata, which provides joint distributions of multiple characteristics but operates at low spatial resolution, making it difficult to identify localized risks; and (2) geographically-aggregated data, which operates at a high spatial resolution but only provides marginal distributions of single characteristics, making it difficult to assess compounding risks. Spatial microsimulation bridges these gaps by integrating the strengths of both data types to create synthetic populations, but its application in U.S. public health research has been constrained by outdated data, limited scope, and insufficient transparency and validation. To address these limitations, we developed and validated a novel synthetic U.S. population for large-scale public health research. Using iterative proportional fitting (IPF), we constrained joint distributions of linking variables from individual-level microdata in the 2022 American Community Survey (ACS) Public Use Microdata Sample (PUMS)—including age, sex, race, ethnicity, income, and housing characteristics—to their known marginal distributions in the 2018–2022 ACS at the census tract level. The resulting synthetic U.S. population of 330 million agents accurately reflected the real U.S. population across key social, economic, infrastructural, and environmental characteristics at a high spatial resolution. Our model demonstrated strong performance against both internal and external datasets, achieving an average Pearson’s R^2 of 0.98, NRSME of 0.11, SAE of 0.03, and E5 of 0.13 across census tracts when benchmarked against the 2018–2022 ACS, and Kullback-Leibler divergence statistics below 0.08 across all variables when benchmarked against the 2020 Decennial Census and the 2020 Residential Energy Consumption Survey, with performance varying by census tract size and diversity and by variable complexity. By building on best practices in

spatial microsimulation, our model offers three key capabilities for future research: (1) enabling the study of spatially explicit joint distributions to uncover compounding risks; (2) facilitating integration with external datasets and models; and (3) gaining access to otherwise difficult-to-assess target variables. Together, these advancements position our synthetic U.S. population as a critical tool for informing the development and deployment of more effective and equitable public health interventions.

2. INTRODUCTION

At the heart of what keeps people healthy are fundamental necessities: clean air, safe water, nutritious foods, healthy buildings, safe neighborhoods, social support systems, and accessible primary care.¹⁴

Yet not everyone has equal access to these necessities, and this inequity cannot be attributed to mere random chance or differences in people's biologies or behaviors. It stems from entrenched systems of oppression that deliberately determine which particular people can be healthy and which particular people should get (and stay) sick.¹⁵ Beyond interpersonal and institutional discrimination that prohibits marginalized communities from accessing quality care, a broader form of historical and ongoing structural discrimination has segregated these communities into neighborhoods characterized by substandard housing, hazardous workplaces, degraded infrastructures, close proximity to toxic sites and polluting facilities, food deserts and fast-food jungles, limited opportunities for quality education and employment, and restricted access to healthcare services and facilities.^{29,30,35–40} These inequities contribute to a wide range of adverse health outcomes disproportionately experienced by marginalized communities, including acute symptoms, infectious disease, chronic conditions, mental illness, violence, injury, and premature death.^{29,30,39–45}

The processes outlined above are rooted in the concept of embodiment, where “[we] literally incorporate, biologically, the world in which we live, including our societal and ecological circumstances” (Krieger, 2014).⁵⁶ **Therefore, the ability to study *who* is most vulnerable and *where* they live is paramount for effectively designing interventions and efficiently dispersing resources.** Central to advancing this effort is the evolution of social and spatial epidemiology. Early social epidemiologic research narrowly focused on assessing single characteristics in relation to health outcomes, largely due to the dominance of reductionist perspectives in research and the limitations of available data and analytical techniques.⁵⁷ However, as the field evolved, it became increasingly clear that alternative approaches that could examine how these single characteristics interact within broader social and spatial contexts to produce health and health inequities were necessary.⁵⁷ The development of spatial epidemiologic research emerged from this recognition and has been propelled by the growing availability of spatial data in conjunction with significant advancements in spatial techniques.^{58–61} Together, these epidemiologic methods have enabled researchers to gain deeper insights into how social, economic, political, and environmental contexts intersect to shape the distribution of health and health inequities across populations.

Despite these advancements, significant methodological challenges remain. These challenges largely stem from the inherent limitations of the two primary types of data available at a large scale: geographically-aggregated data and individual-level microdata.⁵⁴ These formats are often used to protect the privacy of survey respondents, but they limit our ability to accurately identify the most vulnerable populations and meaningfully address their needs.⁵⁴

Geographically-aggregated data is useful for conducting large-scale spatial analyses that identify broad public health trends without compromising privacy. However, this data poses significant limitations when it comes to modeling, interpreting, and intervening in observed relationships. Geographically-aggregated data is generated by summarizing the single characteristics

of survey respondents within specific geographic areas into averages, totals, or proportions, producing marginal distributions of each characteristic. This process smooths over the intricacies of how multiple characteristics interact within individuals (individual variability) and how these interactions differ across geographic areas (spatial variability).⁶² For instance, a neighborhood in which 80% of residents are Black and 80% are low-income does not mean that all Black residents are also low-income, or that the joint effects of race and income on health are consistent across all individuals. These joint effects can also vary significantly within or between different geographic areas, where differences in local social and built environments can modify how these characteristics interact to influence health, even among geographic areas that may share demographic profiles. Additionally, the limited set of variables available in geographically-aggregated data often forces the use of these few variables as proxies for other important but absent target variables. Because proxies are inherently imperfect substitutes for target variables, they inevitably introduce a certain degree of measurement error that can propagate through the analysis, ultimately leading to biased estimates.⁶² These oversimplifications worsen the consequences of the all-too-common ecological fallacy, where conclusions about individual-level relationships are incorrectly inferred from group-level data, even though the correlations between the same variables at the individual and group levels can differ significantly in magnitude and/or direction.^{62–69} The issues that arise from using geographically-aggregated data can therefore not only obscure, but also distort the nuanced dynamics that underlie real-world health outcomes. This becomes particularly concerning when considering that some of the most widely-used tools for assessing vulnerability—including the Center for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI),⁷⁰ CDC Environmental Justice Index (EJI),⁷¹ White House Climate and Economic Justice Screening Tool (CEJST),⁷² Environmental Protection Agency (EPA) EJScreen,⁷³ Federal Emergency Management Agency (FEMA) Resilience Analysis

and Planning Tool (RAPT)⁷⁴—exclusively rely on geographically-aggregated data and can therefore lead to the development and dissemination of misguided public health interventions.

Conversely, individual-level microdata contains detailed information about survey respondents, addressing many of the limitations inherent in geographically-aggregated data. By capturing the joint distributions of multiple characteristics within individuals, microdata allows researchers to analyze how these characteristics co-occur and interact to influence exposures and outcomes.⁶² This capability is crucial because real-world exposures and outcomes related to health are rarely the result of a single characteristic; rather, they emerge from a complex interplay of social, economic, infrastructural, and environmental characteristics that can heighten vulnerabilities in specific subpopulations.^{15,56,67,75} Returning to our previous example, being low-income and belonging to a minority racialized group can amplify health risks in ways that are not merely additive, but multiplicative, creating double, triple, or even heavier burdens far greater than the sum of their parts.⁷⁶ Moreover, microdata provides access to a broader range of variables, enabling direct analysis of target variables that are often oversimplified or misrepresented in geographically-aggregated data.^{53,77,78} This richness is essential for accurately identifying the most vulnerable subpopulations. However, despite these advantages, microdata comes with its own limitations. Because microdata typically relies on a smaller, “representative” subsample of the population, it may not fully capture the marginal distribution of each characteristic across the entire population.⁶² Another major drawback is that these data are only made available at broader geographic scales than geographically-aggregated data to protect individual respondents’ privacy.⁶² These limitations can make it challenging to use individual-level microdata to produce generalizable findings or capture nuanced public health trends at the local level.

Put simply, a critical gap remains: no single data source situates joint single characteristics within their specific geographic contexts. Given data privacy barriers, researchers face significant

challenges in identifying who is most vulnerable and where they live, making it extremely difficult for policymakers to understand how to design effective, targeted interventions and disperse resources efficiently.

But what if we could overcome these challenges by generating a dataset that combines the strengths of *both* data types? This is the very question that lies at the core of spatial microsimulation. Spatial microsimulation—formally defined as “the creation, analysis, and modeling of individual-level data within geographic zones” (Lovelace & Dumont, 2016)⁵⁴—offers a powerful approach to integrating the granular detail of individual-level microdata with the geographic precision of geographically-aggregated data (**Figure 2**). By effectively “combining the best of both worlds of available data into a single analysis” (Lovelace & Dumont, 2016),⁵⁴ this method produces synthetic spatial microdata (colloquially referred to as a synthetic population), enabling researchers to gain a nuanced understanding of how joint single characteristics interact with and are shaped by their specific geographic contexts—dynamics that are often lost in traditional analyses but are necessary to provide policy-relevant insights.⁵³

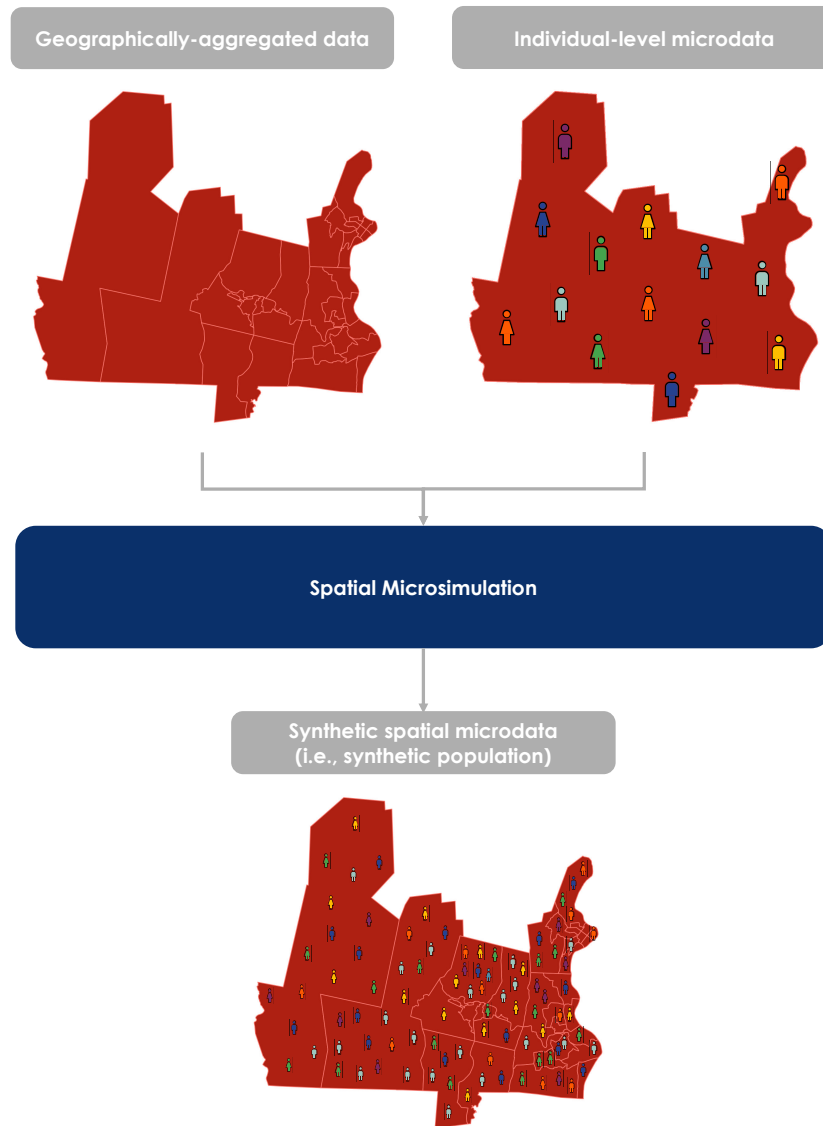


Figure 2. Schematic diagram of spatial microsimulation (adapted from Lovelace & Dumont, 2016).⁵⁴

Although long-established and widely-used in the fields of transportation, urban planning, and economic policy analysis, spatial microsimulation has yet to realize its full potential in public health, particularly in the U.S.⁵⁵ This is surprising given the capacity of synthetic populations to offer detailed insights into how people and places interact to influence health outcomes as well as their utility in simulating future scenarios.⁷⁹ Previous efforts to employ spatial microsimulation in public

health have often been constrained by the use of outdated data, a narrow focus on single exposures or outcomes, and a tendency to restrict analyses to relatively small geographic areas. Furthermore, large-scale U.S. models often suffer from insufficient transparency and inadequate validation, which undermines their reliability for informing policy and practice. Consequently, these models lack the rigor needed to address contemporary and evolving public health challenges.

To address these limitations, we employed a comprehensive approach to develop a census-tract-level synthetic U.S. population that reflects the real U.S. population with respect to a holistic set of social, economic, infrastructural, and environmental characteristics. This approach was designed with the express goal of creating a robust, reliable, and flexible tool that can enable the exploration and analysis of large-scale, complex public health issues and serve as a foundation for integrating other models and/or generating dynamic models. Therefore, this tool is poised to tackle a range of pressing public health questions across the U.S. such as:

1. **Climate justice:** What combination of social, economic, infrastructural, and environmental characteristics determines air conditioning adoption and usage, and how does this intersect with heat stress risks across different census tracts? How can these insights guide the targeted implementation of affordable cooling solutions to protect vulnerable populations from heat-related illness and death?
2. **Environmental justice:** How does the age of residents and the homes they reside in intersect to amplify health risks from lead exposure across different census tracts? How can this information guide the targeted implementation of home lead abatement programs?
3. **Infectious disease control:** What combination of social, economic, infrastructural, and environmental characteristics contribute to vaccine hesitancy, and how does COVID-19 transmission differ across vaccine-hesitant areas? How can localized public health

interventions—including vaccination drives, quarantine measures, and building controls—be tailored to prevent transmission in these areas?

4. **Chronic disease prevention:** Which local factors—such as food deserts, lack of green spaces, and limited walkability—contribute most to the development of chronic diseases like obesity, diabetes, and hypertension in different census tracts? How can these environments be transformed to reduce chronic disease prevalence?
5. **Future scenario projections:** How will population aging patterns in conjunction with domestic migration patterns intersect to reshape the geographic distribution of older adults at risk of social isolation? What preventive support systems can be implemented to mitigate the increasing risk of mental illness in these areas?

In the following sections, we detail the methodologies employed in the creation of this model, present the results of our analysis, and discuss how these findings can inform future public health research and practice.

3. METHODS

Our overall spatial microsimulation approach to generate a census-tract-level synthetic U.S. population involved several key steps that are detailed in subsequent sections, starting with the careful selection of spatial microsimulation model parameters tailored to our research goals and available data. We then processed the data inputs, generated the synthetic U.S. population, and conducted a validation study. All analyses were conducted in R version 4.3.3 (Bioconductor 3.18)⁸⁰ on the Harvard University Faculty of Arts & Sciences Research Computing Cluster.⁸¹

3.1. Spatial microsimulation.

3.1.1. Modeling framework.

To fully appreciate this approach, it is essential to understand the wide range of methods underpinning spatial microsimulation, how they work, and their respective strengths and weaknesses (Figure 3).^{55,82,83} These methods can be categorized as dynamic or static. Dynamic spatial microsimulation generates synthetic spatial microdata that models future life events such as births, deaths, and migration through random sampling or trend forecasting, making it valuable for simulating the long-term impact of demographic shifts or policy changes over time. Static spatial microsimulation, on the other hand, generates synthetic spatial microdata at a specific point in time. This approach can be useful for analyzing the current distribution of attributes and simulating the immediate effects of policy changes, but can also serve as a basis for integrating other models and/or building dynamic models.^{55,82,83}

Static spatial microsimulation can be achieved either through synthetic reconstruction or reweighting methods. Synthetic reconstruction is typically employed when suitable microdata is not available, but aggregated data at the small-area level is. This method generates a synthetic population by creating individuals whose characteristics, when aggregated, match their known marginal totals at the small-area level.⁸³ It leverages a combination of techniques like Iterative Proportional Fitting (IPF) to sequentially adjust characteristics to align more closely with known marginal totals^{84–87} and Monte Carlo Simulation to introduce randomness and approximate joint-probability distributions when precise joint data is lacking.^{88,89} However, with the increasing availability of microdata and advancements in computational techniques, reweighting methods have emerged as an alternative to synthetic reconstruction, designed to “allocate individual counts or weights to geographic zones based on a number of linking variables that are shared between microdata and aggregated data”

(Lovelace, 2014)⁶². Probabilistic reweighting uses randomized combinatorial optimization techniques such as hill climbing, genetic algorithms, or simulated annealing to iteratively select and replace individuals in the microdata to best match their combined characteristics with known marginal totals at the small-area level.⁹⁰ The randomness introduced during this process allows for exploring different combinations that may optimize fit but can also lead to varying results with each run. In contrast, deterministic reweighting uses systematic techniques including IPF or generalized regression to iteratively adjust the fractional weights of individuals in the microdata to align their combined characteristics with known marginal totals at the small-area level, guaranteeing the same result with each run.⁵⁵ Although this method distinctly produces fractional weights that may be less practical than integer weights, this issue can be resolved using integerization techniques.^{91,92}

One key technique used in deterministic reweighting methods is IPF. Originally developed by Deming & Stephan in 1940, IPF was not originally intended for simulating populations, but rather for addressing the challenge of updating joint distributions of categorical data when complete counts were not available or possible to collect by combining two or more sets of data.⁸⁴ This technique operates by iteratively scaling the cell counts within a given contingency table—representing the joint distribution of multiple categorical variables—to match their given marginal totals. Over time, however, it has been adapted for spatial microsimulation to allocate individuals into geographic zones by leveraging individual-level microdata that is not spatially resolute and geographically-aggregated data that is spatially resolute.⁹³ In the spatial microsimulation context, IPF is often described as identifying the most plausible configuration of individuals within geographic zones while maintaining consistency with marginal totals across these zones.⁵⁴ The algorithm continues to adjust the weights of individuals across iterations to reduce discrepancies between the current weighted totals and the target marginal totals.⁵⁴ This process continues until convergence is achieved, where further adjustments result in negligible changes, signaling that the synthetic

population has reached a stable configuration.⁵⁴ The computational speed and deterministic nature of IPF ultimately make it an efficient and reliable technique that is often favored over other techniques.^{85–87,94}

We chose to build our synthetic U.S. population with a static deterministic reweighting model using IPF techniques based on our research goals and available data. This model relies on the careful selection of linking variables shared between data inputs as well as an appropriate geographic scale for analysis to ultimately produce a synthetic population that statistically reflects spatially explicit joint distributions in the real population. This approach can be expanded into agent-based or future population simulations, enabling the exploration and analysis of large-scale, complex public health issues—whether they focus on assessing current conditions, predicting future scenarios, or simulating interactions—across diverse geographic contexts.

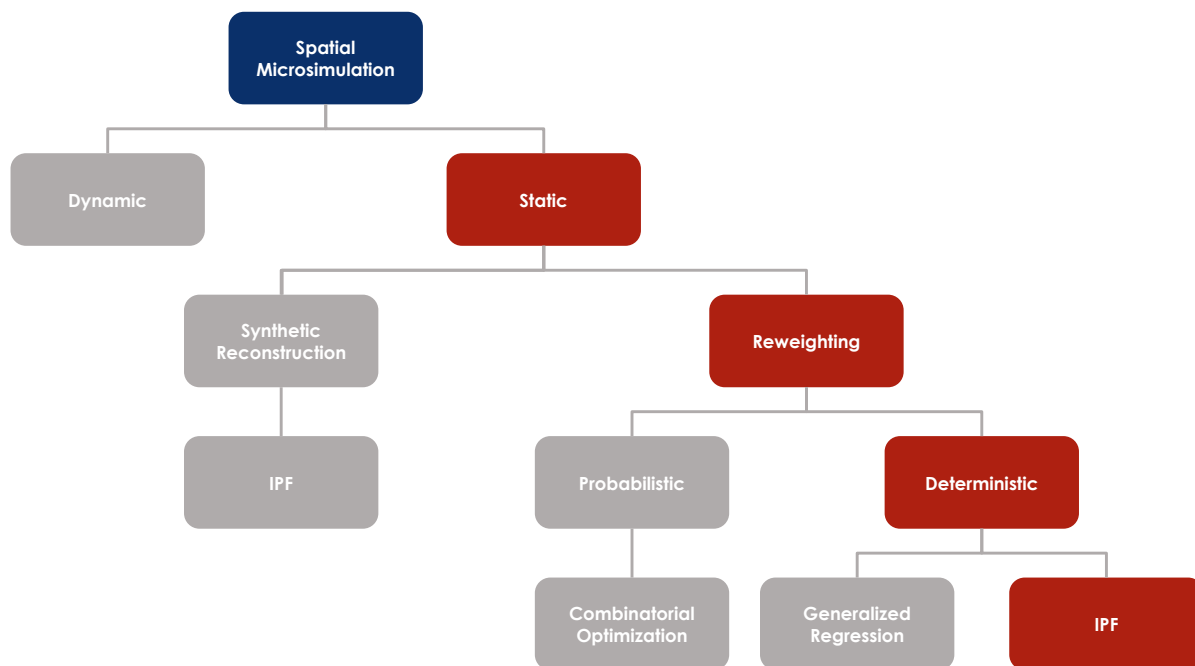


Figure 3. Methodological framework for spatial microsimulation (adapted from Tanton, 2014).⁸³

3.1.2. Linking variables.

Linking variables, which constrain individual-level microdata to geographically-aggregated data, form the backbone of a synthetic population and ultimately determine its relevance for research, policy, and practice. In order for a synthetic population to reliably simulate the real population, these variables must not only be consistent between datasets, but should also be well-distributed, reflective of population diversity, and drivers of population health dynamics.^{95–97} Therefore, in this study, we selected age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel as our potential linking variables. We initially tested model fit using only social variables (age, sex, race, and ethnicity), then with the addition of economic variables (household income), and finally with the addition of environmental variables (housing tenure, housing occupancy, building type, building age, and heating fuel), finding that the inclusion of all variables resulted in an optimal model fit as per the criteria outlined in our internal and external validation studies in later sections. A large body of research supports the decision to utilize this particular suite of linking variables, demonstrating that these are among the key variables that collectively shape where people live and the quality of their environments,^{14,31–33,98,99} how they form relationships and engage within their communities,^{43,44,100–102} and what opportunities and resources are available to them,^{103–105} in addition to how their bodies respond biologically to their surroundings.^{106,107} Together, these factors lie at the core of the complex social, economic, infrastructural, and environmental interactions that drive real-world public health outcomes.^{15,56,75,108} When selected appropriately, they can provide a strong foundation for estimating a wide range of target variables—key exposures and outcomes that are otherwise difficult to assess—across several geographic scales.

3.1.3. Geographic scale.

The decision of which geographic scale to utilize is equally important when constructing a synthetic population because it directly affects its accuracy and utility. Again, the goal of constructing a synthetic population is to introduce spatial granularity to individual-level microdata that is not spatially resolute (operating at the PUMA level in our case) by constraining it to geographically-aggregated data that is spatially resolute (operating at the county, census tract, or census block group levels in our case). Given that geographically-aggregated data is available across several geographic scales, we had the discretion to choose the one that best aligned with our research goals. To make an informed choice, it is essential to understand the differences between them:

- **PUMAs** are defined as non-overlapping regions within each state, containing approximately 100,000 residents and encompassing smaller geographic units like counties, census tracts, and census block groups.¹⁰⁹
- **Counties** are the main political and administrative divisions of states, often used for broad regional analysis.¹⁰⁹
- **Census tracts** are smaller subdivisions of counties that contain around 1,200 to 12,000 people, with a target size of 4,000 people. They are designed to be relatively permanent to enable statistical comparisons across census years and to capture areas with homogenous populations to enable more precise local analysis.¹⁰⁹
- **Census block groups** are even smaller subdivisions of census tracts with populations ranging from 600 to 3,000 people. They consist of clusters of blocks bounded by visible features (streets, roads, streams, tracks) and by nonvisible boundaries (city, township, district, county, property lines) to facilitate small neighborhood analysis.¹⁰⁹

We selected census tract as our geographic scale to capture fine-grained spatial variations in social, economic, infrastructural, and environmental characteristics that are often masked in county-level analyses, while avoiding the issues of data sparsity, noise, and variability that census block-level analyses are often prone to given small sample sizes and disclosure avoidance practices at these levels.^{110–113} Supporting this choice, Krieger et al. demonstrated that the geographic scale at which socioeconomic measures are analyzed significantly influences the observed magnitude of health disparities, with census tracts better reflecting neighborhood-level variations in socioeconomic conditions that shape health outcomes compared to county- or census-block-level analyses.¹¹³ Moreover, the widespread use of census tracts in key datasets like those from the EPA,^{114–116} FEMA¹¹⁷, and National Oceanic and Atmospheric Administration (NOAA)¹¹⁸ facilitates the integration of our synthetic U.S. population with a broad range of existing data sources. Therefore, this choice strengthens both the analytical rigor of our research and the direct applicability of our findings to public health initiatives.

Even though we selected census tract as our geographic scale, because the individual-level microdata operates at the PUMA level, it is crucial to align the census tract and PUMA boundaries. To ensure proper alignment, we created updated national geographic crosswalk tool, as the most recent version from the U.S. Census Bureau did not include the geographic boundary changes in Connecticut (CT) that took effect in early 2022.¹¹⁹ This tool includes both Federal Information Processing Standards (FIPS) codes and corresponding names where applicable. We extracted, cleaned, and merged the special relationship files provided by the U.S. Census Bureau that mapped new PUMA-to-county, county-to-tract, and county-to-block relationships.¹¹⁹ This process enabled us to create a CT-specific crosswalk, which we then integrated into the national crosswalk, resulting in a new crosswalk that reflects all current U.S. geographic boundaries.

3.2. Data.

3.2.1. Geographically-aggregated data.

3.2.1.1. Data collection.

Selecting geographically-aggregated data should be the first consideration when selecting input data. Put plainly: “without good geographical data, there is no way to allocate the individuals to different geographical zones” (Lovelace & Dumont, 2016).⁵⁴

We ultimately utilized geographically-aggregated data from the 2022 5-Year American Community Survey (ACS). The ACS is administered by the U.S. Census Bureau, a federal statistical agency that continuously collects data through detailed long-form questionnaires, which reach approximately 3.5 million addresses each year. This makes the ACS the largest, most comprehensive, and most up-to-date source of social, economic, infrastructural, and environmental characteristics in the United States. We specifically selected the 2022 5-Year ACS for our analysis because it offers a broad temporal scope, capturing the most current trends across the U.S. population while still providing stable estimates of both large and small geographic areas.

The resulting datasets include marginal distributions of key variables that provide a summary of population patterns (**Table 1**). ACS variables utilized as linking variables include age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel. A complete list of variables can be found in the 5-Year 2022 ACS Subject Definitions file.¹²⁰

These data are aggregated at several geographic levels, including counties, census tracts, and census block groups. While the aggregation process maintains individual confidentiality, this hierarchical structure enables the analysis of broad patterns at lower spatial resolutions, detailed patterns at higher spatial resolutions, and intra-spatial patterns. The counties, census tract, and

census block groups used in the 2022 5-Year ACS were based on the new boundaries updated in the 2022 Census year and can be explored through the Census TIGERweb interactive mapping application.¹²¹

Table 1. Geographically-aggregated data – “wide” structure.

Census Tract	Male	Female	<5 Y/O	5-9 Y/O	10-14 Y/O	Rent	Own	...
01001020100	1000	1000	75	100	130	700	1300	
01001020200	800	950	30	55	40	1170	580	
01001020300	3500	3200	120	190	300	3350	3350	
01001020400	2000	2050	100	60	90	2995	1055	

3.2.1.2. Data processing.

To standardize the geographically-aggregated data for use as constraints in IPF, we had to address the varying structures and resulting discrepancies across data tables. We extracted age, sex, race, and ethnicity data from Table S0601; household income data from Table S1901; housing occupancy data from Table S2501; and housing tenure, building type, building age, and heating fuel data from Table DP04 at the census tract level for all 50 states in the U.S. and Washington D.C. from <https://data.census.gov/table>.

We began by converting percentages to counts using the total population for each census tract. We then harmonized the household-level variables (household income, housing occupancy, housing tenure, building type, building age, and heating fuel) with individual-level variables (age, sex, race, and ethnicity). To do this, we estimated the average household size by dividing the total population by the total number of households for each census tract and applied this average to each variable subcategory. Finally, we merged all datasets by census tract to create a single constraints dataset.

We performed further refinements to address inconsistencies where the variable totals in the ACS tables did not match the population totals. These discrepancies likely arose due to differences in sampling variability, rounding, and missing data across the ACS tables. We explored two methods to address these discrepancies: one involved adjusting the totals by adding or subtracting from the most common subcategory, while the other involved proportionally adjusting all subcategories. We ultimately selected proportional adjustment because this strategy best maintained the specific distributions of variable subcategories and yielded a synthetic U.S. population that more closely aligned with internal and external sources as per the validation study described in later sections. To implement this, we first calculated a variable total by summing the values of all its subcategories. We then compared this variable total to the population total and proportionally distributed this difference across the subcategories. Specifically, we divided this difference by the variable total to get an adjustment factor, multiplied the adjustment factor by each subcategory value, and added or subtracted the result back to each subcategory value to get an adjusted subcategory value. We also included safeguards to ensure that no adjusted subcategory value fell to or below zero, except where originally zero. This process was systematically applied to ensure that all variable totals aligned with the population totals for each census tract while maintaining data integrity.

After completing these data standardization and data refinement steps, we finally turned our attention to consolidating variable subcategories. We consolidated specific subcategories of racialized groups, building types, building ages, and heating fuels. This process was crucial because the presence of too many cells with zero or very low counts could hinder the IPF algorithm's ability to iteratively adjust the data to fit the marginal totals. By addressing data sparsity, we created a more stable data matrix, which improved the IPF algorithm's ability to reliably converge and produce accurate results. However, we were careful not to over-consolidate, as doing so could obscure important variability present in the real population and diminish the analytical value of the synthetic

U.S. population. Striking this balance was critical in creating a synthetic U.S. population that was both computationally manageable and statistically representative of population health dynamics, ensuring it could serve as a practical yet powerful research tool.

3.2.2. Individual-level microdata.

3.2.2.1. Data collection.

However, while geographically-aggregated data forms a solid foundation for analyzing population patterns across zones, it omits several important variables and cannot capture variability within those zones.

To address these limitations, we utilized data on individuals and households from the 1-Year 2022 American Community Survey (ACS) Public Use Microdata Sample (PUMS). The PUMS dataset is a representative subset of the larger ACS dataset that provides anonymized individual- and household-level microdata on about 1% of the U.S. population. We chose the 1-Year 2022 ACS PUMS for our analysis because it offers the most current information on joint distributions within the U.S. population while still maintaining geographic consistency given the most recent changes made to U.S. Census boundaries.

The resulting dataset maintains linkages between individuals and households to allow for flexibility in analytical unit and includes a wide range of variables that capture joint distributions of multiple characteristics across the U.S. population (**Table 2**). The ACS PUMS provides linking variables that are shared with the standard ACS including age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel; it also provides unique variables that can be used as target variables following the construction of the synthetic U.S. population such as household relationships and family structures, language spoken

and nativity status, housing, utility, and energy costs, occupation, industry, and employment details, insurance status and public assistance participation, and commuting patterns and transportation modes. A complete list of variables can be found in the 1-Year 2022 ACS PUMS Data Dictionary file.¹²²

These data are available at the Public Use Microdata Area (PUMA) level, which enables the release of detailed individual- and household-level microdata while preserving data privacy. The PUMAs used in the 2022 1-Year ACS PUMS were based on boundaries established in the 2020 Decennial Census. These PUMA boundaries can be explored through the Census TIGERweb interactive mapping application.¹²¹

Table 2. Individual-level microdata – “long” structure.

Person ID	PUMA	Sex	Age	Race	Household Income	Housing Tenure	Building Type	...
1	1700	Male	28	White	35000	Rent	50+ APT	
2	2701	Female	82	Black	65000	Own - full	SFH	
3	2500	Female	5	Asian	155000	Occupied no rent	2 APT	
4	11000	Male	44	Native	70000	Own - mortgage	SFH	

3.2.2.2. Data processing.

To prepare the individual-level microdata for Iterative Proportional Fitting (IPF), we extracted age, sex, race, and ethnicity information from the individual-level files and household income, housing tenure, housing occupancy, building type, building age, and heating fuel information from the household-level files available at <https://www2.census.gov/programs->

[surveys/acs/data/pums/2022/1-Year/](#). We then merged these datasets by unique ID to link individuals to their corresponding households.

IPF relies on having a complete dataset to function effectively. To achieve this, we first assessed the extent and nature of missingness, finding less than 5% missingness across all variables with no discernable pattern. We then tested two imputation strategies: mode imputation, which fills in missing data with the most common variable subcategory, and proportional imputation, which fills in missing data based on the observed distribution of variable subcategories. Each strategy was applied at three different scales (across the entire sample, within PUMAs, and within age groups within PUMAs) using two data structures (raw and banded), resulting in a total of twelve distinct tests. Consistent with our approach for handling the constraints data, we ultimately chose proportional imputation within age groups and PUMAs using raw data because this strategy best maintained the specific distributions of variable subcategories and yielded a synthetic U.S. population that more closely aligned with internal and external sources as per the validation study described below.

Equally crucial for IPF, however, is having complete representation of all variable subcategories within each PUMA. Because proportional imputation maintained the original distribution of variable subcategories observed in the PUMS data, any variable subcategories not present in the PUMS data for a specific PUMA remained missing after the imputation process. This could occur even if certain variable subcategories were expected to be present in the population residing in that specific PUMA due to the sampling limitations of PUMS which may underrepresent hard-to-reach populations. Without these variable subcategories, the IPF process cannot properly constrain the individual-level microdata to the geographically-aggregated data, leading to potential inaccuracies in the synthetic U.S. population. To address this issue, we sampled two individuals from a neighboring PUMA who had the missing attributes and added them to the PUMA with incomplete

representation, ensuring that the synthetic U.S. population accurately represented all variable subcategories.

Finally, while IPF requires banding variables to align with the structure of the geographically-aggregated data, we chose to retain the raw variables as well. This careful balance of imputation, sampling, and variable retention allowed us to create a synthetic U.S. population that is not only reliable but also robust enough to offer detailed insights in further analysis (**Table 3**).

Table 3. Comparison of variable sub-categories across input datasets.

Variable	Individual-level microdata		Geographically-aggregated data	
	Units	Source	Units	Source
Geography	Public Use Microdata Area (PUMA)	2022 1-Year ACS PUMS	Census Tract	2022 5-Year ACS
Age	Years	2022 1-Year ACS PUMS - Person	< 5 Years 5-17 Years 18-24 Years 25-44 Years 45-54 Years 55-64 Years > 75 Years	2022 5-Year ACS S0601
Sex	Male Female	2022 1-Year ACS PUMS - Person	Male Female	2022 5-Year ACS S0601
Race	White Asian Black American Indian Native Alaskan Native Hawaiian Pacific Islander Some other race alone Two or more races	2022 1-Year ACS PUMS - Person	White Asian Black Native (American Indian, Native Alaskan, Native Hawaiian, or Pacific Islander) Other (some other race alone or two or more races)	2022 5-Year ACS S0601
Ethnicity	Hispanic Non-Hispanic	2022 1-Year ACS PUMS - Person	Hispanic Non-Hispanic	2022 5-Year ACS S0601
Household Income	Dollars	2022 1-Year ACS PUMS - Household	< \$10,000 \$10,000-\$14,999 \$15,000-\$24,999 \$25,000-\$34,999	2022 5-Year ACS S1901

			\$35,000-\$49,999 \$50,000-\$74,999 \$75,000-\$99,999 \$100,000-\$149,999 \$150,000-\$199,999 > \$200,000	
Housing Tenure	Owned with mortgage or loan Owned free and clear Rented Occupied without payment of rent	2022 1-Year ACS PUMS - Household	Own Rent	2022 5-Year ACS DP04
Housing Occupancy	Number of people	2022 1-Year ACS PUMS - Household	1 Person 2 People 3 People > 4 People	2022 5-Year ACS S2501
Building Age	< 1940 1940-1949 1950-1959 1960-1969 1970-1979 1980-1989 1990-1999 2000-2009 2010-2019 2020 2021 2022	2022 1-Year ACS PUMS - Household	< 1950 (< 1940, 1940-1949) 1950-1979 (1950, 1959, 1960-1969, 1970-1979) 1980-1999 (1980-1989, 1990-1999) > 2000 (2000-2009, 2010-2019, >2020)	2022 5-Year ACS DP04
Building Type	Single family home – detached Single family home – attached 2 apartments 3-4 apartments 5-9 apartments 10-19 apartments 20-49 apartments > 50 apartments Mobile home or trailer Boat, RV, or van	2022 1-Year ACS PUMS - Household	Single family home – detached Single family home – attached 2 apartments 3-4 apartments 5-9 apartments 10-19 apartments > 20 apartments Mobile (mobile home, trailer, boat, RV, or van)	2022 5-Year ACS DP04
Heating Fuel	Coal or coke Fuel oil or kerosine Wood Utility gas	2022 1-Year ACS PUMS - Household	Coal or oil (coal, coke, fuel oil, kerosine) Biomass	2022 5-Year ACS DP04

	Bottled, tank, or LP gas Electricity Solar energy Other fuel No fuel used		Natural gas (utility gas, bottled, tank, or LP gas) Electric or renewables (electricity, solar energy) Other fuel No fuel used	
--	--	--	--	--

3.3. Analysis.

After preparing both the geographically-aggregated data and individual-level microdata, we turned our attention to the implementation phase. We based our implementation process on the IPF code outlined in the *Spatial Microsimulation with R* by Lovelace & Dumont in 2016.⁵⁴ To manage the large dataset, we split the PUMAs into batches of 100. Each batch was processed using a custom function that looped through the individual PUMAs, executing the following major steps:

1. **Categorization:** We first categorized the individual-level microdata by converting categorical variables into binary dummy variables (0/1). This resulted in a binary matrix where each subcategory of a variable was represented as a separate column. We then re-ordered the columns of this matrix to align with their corresponding columns in the geographically-aggregated data. This alignment was crucial for ensuring that the individual-level microdata could be constrained to the geographically-aggregated data in subsequent steps.
2. **Initialization:** Following categorization, we initialized the process by assigning the initial weights for each individual to 1 so that each individual started with equal influence.
3. **Iteration:** We then employed the Iterative Proportional Fitting (IPF) algorithm to iteratively adjust the weights of the individuals within each PUMA using the `ipfp` package in R.

Specifically, for each individual i within each zone z , the weight $w(i, z, t)$ was updated at each iteration t using the following formula (Fienberg, 1970):

$$w(i, z, t + 1) = w(i, z, t) \times \frac{cons_t(z, c, ind(i, c))}{\sum_{j=1}^{n_{ind}} w(j, z, t) \times I(ind(j, c) = ind(i, c))}$$

Here, the numerator represents the marginal total for zone z and subcategory c while the denominator sums the current weights of all individuals in the same subcategory. This iterative adjustment conformed the individual-level microdata as closely as possible to the known constraints from the geographically-aggregated data. To maintain numerical stability, we added a small regularization term (1e-10) to the constraints to prevent zero values, and replaced any resulting NA weights with a small positive value (1e-10).

4. **Convergence:** As the IPF process continued, we monitored its convergence by evaluating how well the iteratively adjusted weights aligned with the known constraints. We assessed convergence at three distinct stages—after 10, 20, and 50 iterations—as part of the validation study described in later sections. Ultimately, we allowed it to run for a maximum of 50 iterations, although it was designed to terminate earlier if convergence was achieved before reaching this limit. Convergence was considered successful when the differences between the iteratively adjusted weights and the known constraints were minimized to an acceptable tolerance threshold of 1.5e-8. If convergence was not achieved within 50 iterations, the process was automatically halted, using the weights from the final iteration as the best available adjustment.
5. **Integerization:** We converted the resulting fractional weights from the IPF process into integers through the pseudorandom Truncate, Replicate, Sample (TRS) method.⁹¹ This involved truncating fractional weights to the nearest integer, calculating the residuals (fractions left after truncation) and determining the deficit (the difference between the

desired total and the current total), and randomly distributing these residuals across the population to meet the deficit. The TRS method ensured that the final weights were integers while maintaining the proportionality of the original weights.

6. **Expansion:** Finally, we expanded the integerized data by replicating individuals according to their integer weights. This expansion step generated the final synthetic U.S. population dataset for each PUMA within each batch.

Each batch of processed data was saved as an intermediate file and later combined into a single dataset. This dataset was then merged with a crosswalk file that linked PUMAs to their corresponding census tracts, achieving accurate geographic alignment. The final synthetic U.S. population dataset was saved for further analysis.

3.4. Validation study.

In order to select the model that best fit the synthetic U.S. population to the real U.S. population at the census tract level, we conducted an extensive validation study that included both internal and external validation steps. As described in prior sections, we tested the performance of our model with various linking variables, adjustment strategies, imputation strategies, and number of iterations.

3.4.1. Internal validation.

Internal validation is the “process of comparing the model’s output against data that is internal to the model itself” (Lovelace & Dumont, 2016)⁵⁴ to test its internal consistency. We applied this by converting the synthetic spatial microdata to the same marginal structure as the geographically-aggregated data and comparing the two both visually and using a suite of goodness-

of-fit metrics. These metrics included Pearsons r , root mean squared error (RMSE), normalized root mean squared error (NRSME), total absolute error (TAE), standardized absolute error (SAE), modified Z-score, and the proportion of values deviating from expectations $E > 5\%$ (E5). Their definitions, ranges, and thresholds for performance are detailed in **Table 4** and a more in-depth discussion of their strengths and limitations can be found in Lovelace et al. (2015).¹²³

Table 4. Goodness-of-fit metrics for internal validation.

Metric	Definition	Range	Performance Thresholds
Pearsons r	Measures the strength and direction of the linear relationship between simulated and observed values	-1 to 1	Good: $r > 0.8$ Moderate: $0.5 < r \leq 0.8$ Poor: $r \leq 0.5$
RMSE	Measures the average magnitude of error between simulated and observed values	0 to ∞	Good: close to 0
NRMSE	RMSE divided by the range of observed values (expressed as a percentage)	0 to ∞	Good: $\text{NRMSE} < 0.20$ Moderate: $20\% \leq \text{NRMSE} < 0.30$ Poor: $\text{NRMSE} \geq 0.30$
TAE	Measures the sum of absolute differences between simulated and observed values across all categories	0 to ∞	Good: close to 0
SAE	TAE normalized by the expected variation in the data	0 to ∞	Good: close to 0 (Clarke & Madden (2001) used an error threshold of 80% of areas with $\text{SAE} \leq 0.2$; Smith et al (2007) used $\text{SAE} \leq 0.1$)
Modified Z-score	Measures how many standard deviations an observation is from the mean, adjusted for small sample sizes	$-\infty$ to ∞	--
E5	Measures the proportion of predicted values that deviate from the observed values by more than 5%	0% to 100%	Good: close to 0 (Lovelace et al. (2015) achieved an E5 of 13.2%)

To thoroughly assess the model's internal consistency, we calculated each metric at three levels: each census tract, averaged across all census tracts, and across the entire dataset. This multi-

level approach enabled us to capture the model’s performance at different scales of analysis. Given that some metrics are more robust to variations in population size, we primarily relied on Pearson’s r , NRSME, SAE, and E5 for cross-comparability between different model options and chose the model with the highest Pearson’s r and lowest NRSME, SAE, and E5 as our final model.¹²³ This allowed for a more consistent evaluation of the model’s performance across modeling scenarios.

3.4.2. External validation.

Complementary to internal validation, external validation involves “the process of comparing model results with external data... [which] relies on good ‘ground-truth data’ and can take place at either the individual or the aggregate level” (Lovelace & Dumont, 2016).⁵⁴ In our study, we conducted external validation at the aggregate level by comparing our synthetic spatial microdata against selected ground-truth data sources at their most granular geographic levels available. Specifically, we used the 2020 Decennial Census for validating age, sex, race, and ethnicity distributions at the census tract level, the 2020 Internal Revenue Service (IRS) Tax Year data for household income distributions at the county level, and the 2020 Residential Energy Consumption Survey (RECS) for housing tenure, housing occupancy, building type, building age, and heating fuel distributions at the state level. We first visually compared the simulated and observed distributions of each variable under each modeling scenario. We then employed the Kullback-Leibler (KL) divergence statistic to quantitatively assess the alignment between the simulated and observed distributions of each variable under each modeling scenario using following equation:

$$D_{KL}(P \parallel Q) = \sum_{x \in X} P(x) \log \left(\frac{Q(x)}{P(x)} \right)$$

Here, the KL statistic quantifies the information loss when a simulated probability distribution Q approximates a “true” probability distribution P , with KL statistics closer to 0

indicating minimal information loss and a more accurate representation of the true population. We thus selected the model with the lowest KL statistics as our final model.

4. RESULTS.

We employed a static deterministic reweighting approach to spatial microsimulation to generate an up-to-date synthetic U.S. population. Through the careful selection and validation of model inputs, model parameters, and analytical techniques, we ensured that our synthetic U.S. population statistically reflects the real U.S. population across age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel at the census tract level. An illustrative example of the resulting data structure can be seen in **Table 5**.

Table 5. Synthetic U.S. population dataset – “long” structure

Person ID	Census Tract ID	Sex	Age	Race	Household Income	Housing Tenure	Building Type	...
000000 001	019582 03849	Male	28	White	35000	Rent	20+ APT	
000000 002	019582 03850	Female	82	Black	65000	Own	SFH	
000000 003	019583 00000	Female	5	Asian	155000	Own	2 APT	
000000 004	019583 00001	Male	44	Native	70000	Own	SFH	
000000 005	020010 00001	Male	18	Black	0	Rent	20+ APT	
...								
332672 026	527005 00004	Female	64	Native	90000	Own	SFH	

4.1. Synthetic U.S. population characteristics.

Our final synthetic U.S. population consisted of 330,587,563 agents, each uniquely characterized to reflect the diversity of the U.S. population. The mean age was 39 years, reflecting a diverse age distribution, with an even split between males and females. The majority of the population identified as White (66%), followed by Black (12%), Asian (5.8%), and Native (1%), with 15% identifying as “Other.” Additionally, 19% of the population identified as Hispanic. The median household income was \$75,500.

Housing characteristics showed considerable diversity. While 34% of the population rented their homes, the majority were homeowners, with 41% holding a mortgage and 23% owning their homes outright. The housing stock was predominantly composed of single-family detached homes (63%), with the remainder in attached homes, multi-family units, or mobile homes, most of which were constructed between 1950 and 1999. Natural gas was the primary heating source for 52% of households, followed by electricity at 40%, with a small percentage relying on other sources or no fuel at all. All descriptive statistics for both raw and banded variables can be found in **Table 6**.

Table 6. Descriptive statistics of the synthetic U.S. population dataset.

Characteristic	N = 330,587,563 ¹
Age (Years)	39 (23) 38 [0, 97]
Age (Banded)	
>5 Years	18,959,774 (5.7%)
5-17 Years	54,127,719 (16%)
18-24 Years	30,867,657 (9.3%)
25-44 Years	88,091,012 (27%)
45-54 Years	41,056,071 (12%)

Characteristic	N = 330,587,563¹
55-64 Years	42,602,111 (13%)
65-74 years	32,327,853 (9.8%)
<75 Years	22,555,366 (6.8%)
Sex	
Female	166,805,584 (50%)
Male	163,781,979 (50%)
Race	
Other	49,274,590 (15%)
Native	3,376,045 (1.0%)
Black	41,053,109 (12%)
Asian	19,070,789 (5.8%)
White	217,813,030 (66%)
Ethnicity	
Hispanic	61,573,472 (19%)
Not Hispanic	269,014,091 (81%)
Household Income (\$)	
	105,671 (115,704)
	75,300 [-17,000, 2,481,200]
Household Income (Banded)	
<\$10000	16,028,013 (4.8%)
\$10000-\$14999	12,133,302 (3.7%)
\$15000-\$24999	22,795,785 (6.9%)
\$25000-\$34999	24,149,986 (7.3%)
\$35000-\$49999	35,025,413 (11%)
\$50000-\$74999	52,915,182 (16%)
\$75000-\$99999	42,359,016 (13%)
\$100000-\$149999	57,097,066 (17%)

Characteristic	N = 330,587,563¹
\$150000-\$199999	29,701,888 (9.0%)
>\$200000	38,381,912 (12%)
Housing Tenure	
Rent	113,557,401 (34%)
Occupied - No Rent	5,239,921 (1.6%)
Own - Mortgage	135,530,410 (41%)
Own - Full	76,259,831 (23%)
Housing Tenure (Banded)	
Rent	113,557,401 (34%)
Own	217,030,162 (66%)
Housing Occupancy	2.61 (1.62) 2.00 [1.00, 20.00]
Housing Occupancy (Banded)	
1 Occupant	87,941,423 (27%)
2 Occupants	109,897,089 (33%)
3 Occupants	52,587,938 (16%)
>3 Occupants	80,161,113 (24%)
Building Type	
Boat/RV/Van	422,543 (0.1%)
Mobile Home	18,231,592 (5.5%)
50+ Apartments	17,722,163 (5.4%)
20-49 Apartments	11,170,240 (3.4%)
10-19 Apartments	13,110,921 (4.0%)
5-9 Apartments	14,936,175 (4.5%)
3-4 Apartments	13,977,509 (4.2%)
2 Apartments	11,300,350 (3.4%)

Characteristic	N = 330,587,563¹
Single Family Attached	20,473,384 (6.2%)
Single Family Detached	209,242,686 (63%)
Building Type (Banded)	
Mobile Home	18,654,135 (5.6%)
20+ Apartments	28,892,403 (8.7%)
5-19 Apartments	28,047,096 (8.5%)
2-4 Apartments	25,277,859 (7.6%)
Attached Single Family Home	20,473,384 (6.2%)
Detached Single Family Home	209,242,686 (63%)
Year Built	
<1940	38,344,528 (12%)
1940-1949	14,627,417 (4.4%)
1950-1959	33,249,208 (10%)
1960-1969	33,329,736 (10%)
1970-1979	47,899,418 (14%)
1980-1989	44,286,086 (13%)
1990-1999	42,881,310 (13%)
2000-2009	41,616,779 (13%)
2010-2019	29,132,110 (8.8%)
2020	2,936,531 (0.9%)
2021	1,868,609 (0.6%)
2022	415,831 (0.1%)
Year Built (Banded)	
<1950	52,971,945 (16%)
1950-1979	114,478,362 (35%)
1980-1999	87,167,396 (26%)

Characteristic	N = 330,587,563 ¹
>2000	75,969,860 (23%)
Heating Fuel	
Coal or Coke	242,864 (<0.1%)
Fuel Oil, Kerosine, Etc	13,979,441 (4.2%)
Wood	4,941,268 (1.5%)
Bottled, Tank, or LP gas	18,048,963 (5.5%)
Utility Gas	155,117,630 (47%)
Electricity	131,396,970 (40%)
Solar Energy	1,132,551 (0.3%)
Other Fuel	1,568,080 (0.5%)
No Fuel	4,159,796 (1.3%)
Heating Fuel (Banded)	
Oil or Coal	14,222,305 (4.3%)
Biomass	4,941,268 (1.5%)
Natural Gas	173,166,593 (52%)
Electricity or Renewables	132,529,521 (40%)
Other Fuel	1,568,080 (0.5%)
No Fuel	4,159,796 (1.3%)

¹n (%); Mean (SD)
Median [Range]

4.2. Validation study.

We used several data sources to select and validate the robustness of our final model. For internal validation, we relied on the 2022 ACS data to assess the model's alignment across constraint variables, choosing the model with the highest Pearsons R and lowest NRSME, SAE, and E5 as our

final model. For external validation, we used 2020 Decennial Census data for age, sex, race, and ethnicity at the census tract level, 2022 IRS data for income at the county level, and 2020 RECS data for all housing characteristics at the state level (representing the finest spatial scale available for each data source), choosing the model with the lowest KL divergence statistics as our final model. Importantly, the model selected based on internal validation was the same as the model selected through external validation, eliminating the need to reconcile different optimal solutions. Across all variables, the alignment was generally closer to the 2022 ACS than to external sources, which was expected given that the synthetic U.S. population was constructed based on marginal distributions from the 2022 ACS and was not subject to discrepancies in timing, sampling, or data collection methods between different population surveys.

4.2.1. Visual Validation.

We present visual validation at the county level for age, sex, race, ethnicity, and household income for several reasons: it allows for easier interpretation, accounts for within-county population shifts since 2020, and provides a practical comparison for all socioeconomic variables given data availability. The simulated distributions from the synthetic U.S. population closely mirror the observed distributions of age, sex, race, and ethnicity from the 2022 ACS and 2020 Decennial Census (**Figures 4.1-4.4**). Validating income presented unique challenges due to the inherent difficulty of measuring income itself.¹²⁴ Income tends to fluctuate over time, varies by household size, and is influenced by geographic factors, making it particularly difficult to measure.¹²⁴ As such, we validated our model's household income estimates against the 2020 IRS individual income data, which was the most comparable data available at a fine spatial scale. However, because our model captures household income while the IRS data reflect individual income, the simulated values are

naturally skewed higher than the observed values. Despite this, the overall shape of the two distributions is similar, demonstrating alignment in the relative income patterns (**Figure 4.5**).

Examining these trends across different counties reveals that the model’s performance varies with population size and diversity. In larger, more diverse counties, the model performs exceptionally well. However, this performance diminishes slightly in smaller and less diverse counties. For instance, in Loving County, Texas—where the population size is only 64 people, with most residents concentrated at the younger and older extremes of the age distribution and predominantly identifying as Non-Hispanic White—deviations in the alignment of certain age and racialized subgroups are observed. This difference likely arises from the IPF algorithm’s reliance on rich datasets to capture the full complexity of real-world interactions. In smaller, less diverse populations, the algorithm has fewer data points and less variability for precise weight adjustments, making it more susceptible to converging on local optima—a situation where the algorithm finds a solution that appears optimal within the limited population but may not be globally optimal across the entire population—and over- or under-representing certain subcategories.

Given the lack of data availability on housing characteristics at finer spatial scales, we conducted visual validation at the state level for housing tenure, housing occupancy, building type, building age, and heating fuel (**Figures 4.6-4.10**). The alignment between the synthetic U.S. population and both the 2022 ACS and the 2020 RECS is generally strong across all housing variables and states. Notable differences are observed in Hawaii and Florida, where the RECS data indicate a higher proportion of households reporting no fuel for heating, whereas the synthetic U.S. population and ACS data suggest a greater reliance on electric heating. These discrepancies may result from the reduced heating needs in these warmer climates, which could have influenced how respondents reported their heating sources.



Figure 4.1. Comparison of age distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in (a) the most populous county in each state and (b) the least populous county in each state.



Figure 4.2. Comparison of sex distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in (a) the most populous county in each state and (b) the least populous county in each state.



Figure 4.3. Comparison of race distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in (a) the most populous county in each state and (b) the least populous county in each state.



Figure 4.4. Comparison of ethnicity distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in (a) the most populous county in each state and (b) the least populous county in each state.



Figure 4.5. Comparison of income distributions across the synthetic U.S. population, 2022 ACS data, and 2020 Decennial Census data in (a) the most populous county in each state and (b) the least populous county in each state.

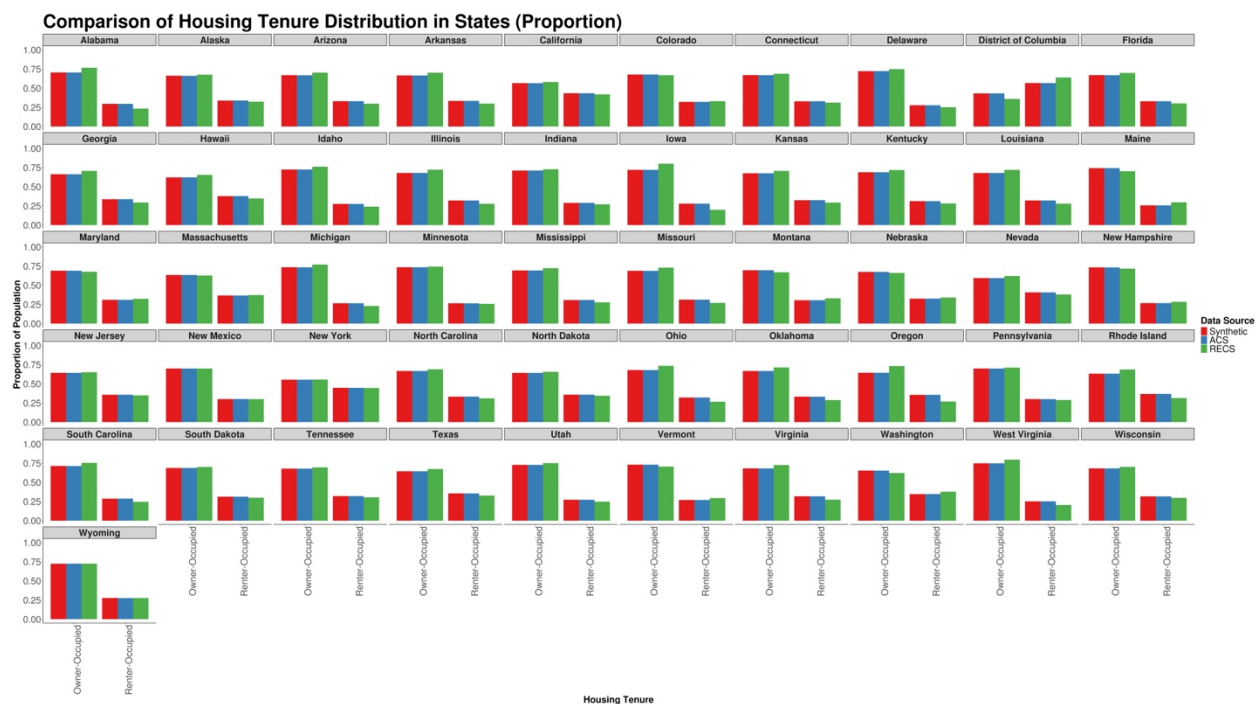


Figure 4.6. Comparison of housing tenure distributions across the synthetic U.S. population, 2022 ACS data, and 2020 RECS data in each state.



Figure 4.7. Comparison of housing occupancy distributions across the synthetic U.S. population, 2022 ACS data, and 2020 RECS data in each state.

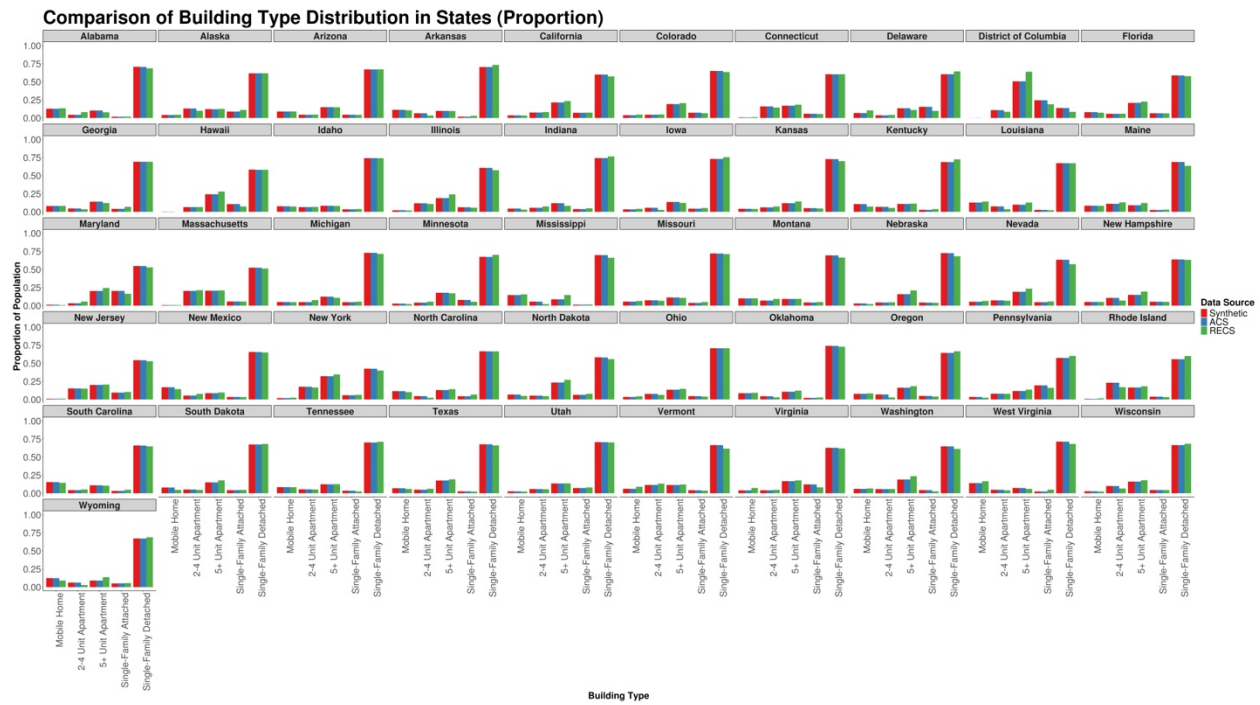


Figure 4.8. Comparison of building type distributions across the synthetic U.S. population, 2022 ACS data, and 2020 RECS data in each state.



Figure 4.9. Comparison of building age distributions across the synthetic U.S. population, 2022 ACS data, and 2020 RECS data in each state.



Figure 4.10. Comparison of heating fuel type distributions across the synthetic U.S. population, 2022 ACS data, and 2020 RECS data in each state.

4.2.2. Internal validation.

The internal validation metrics (**Table 7**) demonstrate that the synthetic U.S. population produced by the final model (predicted) effectively converged to the constraint data (observed) at the census tract level. The Pearson’s R value—which averaged 0.991 with a median of 0.998—indicates a strong correlation between the predicted and observed data, with 99.64% of tracts exhibiting a correlation above 0.80. Similarly, 96.73% of tracts had an SAE below 0.10, suggesting minimal deviation in the predicted values relative to observed values.

The NRMSE and E5 metrics, however, revealed different aspects of the model’s performance given the NRSME’s sensitivity to the magnitude of errors and the E5’s sensitivity to the frequency of errors. The NRSME had a mean of 0.109 but a median of 0.077, implicating the presence of a few tracts with larger errors. In contrast, the E5 showed that a mean of 13.06% and

median of 12.24% of predicted values deviated more than 5% from the observed values, highlighting how often these larger errors occurred across tracts. While these metrics suggest areas where the model’s performance could be impacted by outliers, the strong performance indicated by Pearson’s R and SAE confirms the robustness of the synthetic U.S. population.

Table 7. Internal validation results.

Metric	Pearson’s R	NRSME	SAE	E5
Mean	0.991	0.109	0.032	0.131
Median	0.998	0.077	0.024	0.122
[Min, Max]	[0.299, 0.999]	[0.002, 2.121]	[0.001, 1.105]	[0, 0.796]

As confirmed by **Figure 5**, population size does appear to influence model performance, where smaller tracts tend to exhibit more variability and lower correlations while larger tracts maintain consistently high correlations. A linear regression of Pearson’s R against total population size revealed a statistically significant relationship ($p < 0.0001$), with a small positive effect (intercept = $9.891e-01$, coefficient = $4.283e-07$), suggesting that other factors such as population diversity may play a more substantial role in driving model performance.

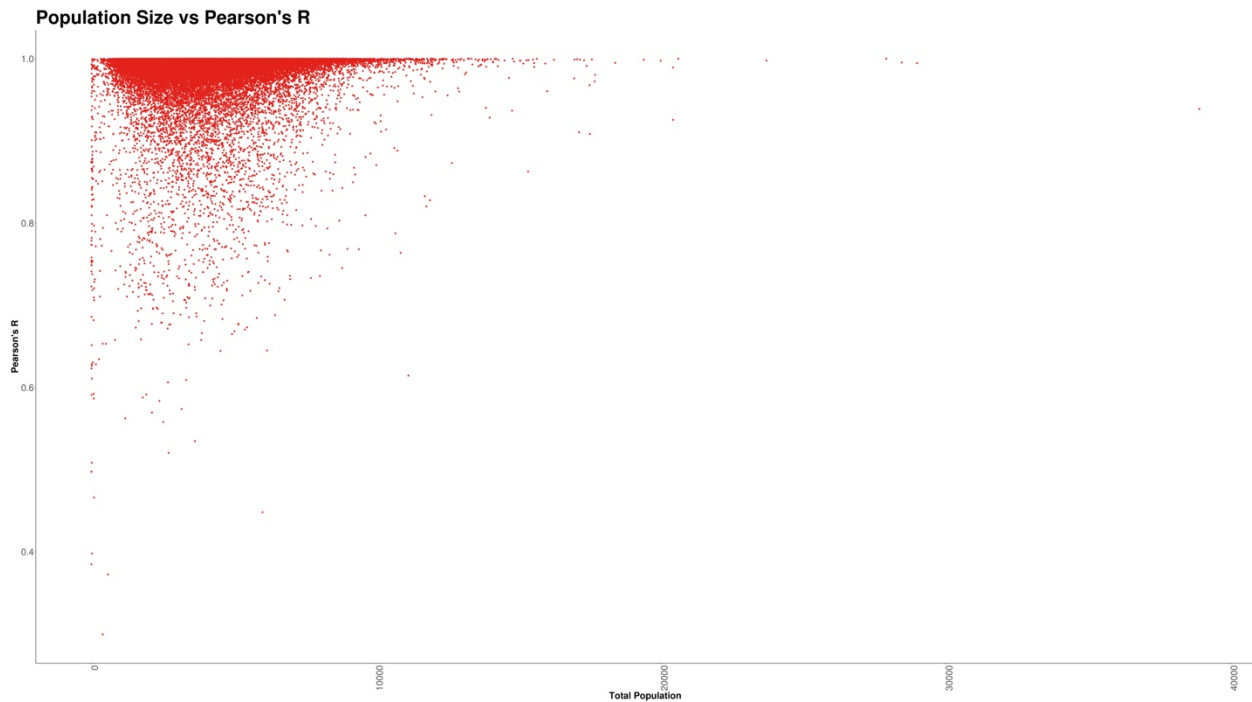


Figure 5. Pearson’s R versus population size by census tract.

4.2.3. External validation.

To further validate the predictive accuracy of the synthetic U.S. population, we compared it against external data sources using the KL divergence statistic, which measures the information loss when approximating a “true” distribution with a simulated one. For this analysis, we used the 2020 Decennial Census data as the “true” distributions for demographic variables at the census tract level and the 2020 RECS data as the “true” distributions for housing variables at the state level. While some divergence is to be expected due to differences in data sources—particularly since our synthetic U.S. population is based on 2022 data whereas the external data were collected two years prior—the KL divergence statistics across all variables were still heavily skewed towards zero, indicating very little information loss across the majority of census tracts and states (**Figures 6a-i**).

Among the demographic variables, the KL divergence statistics for sex and ethnicity were the lowest, with means of 0.004 and 0.005, respectively. In contrast, age and race exhibited slightly higher KL divergence statistics, with means of 0.080 for age and 0.050 for race. These higher values likely reflect the added complexity of replicating finer-grained categories, such as specific age and racialized subgroups, which have more variability and therefore present a greater challenge for the model to capture accurately.

We observed a similar pattern in the housing variables, where those with finer-grained categories also exhibited slightly higher KL divergence statistics. However, all KL divergence statistics across all housing variables were extremely low, each below a mean of 0.015. This is likely due to the validation being conducted at the state level, where the aggregation of data tends to smooth out local variations, naturally leading to lower KL values.

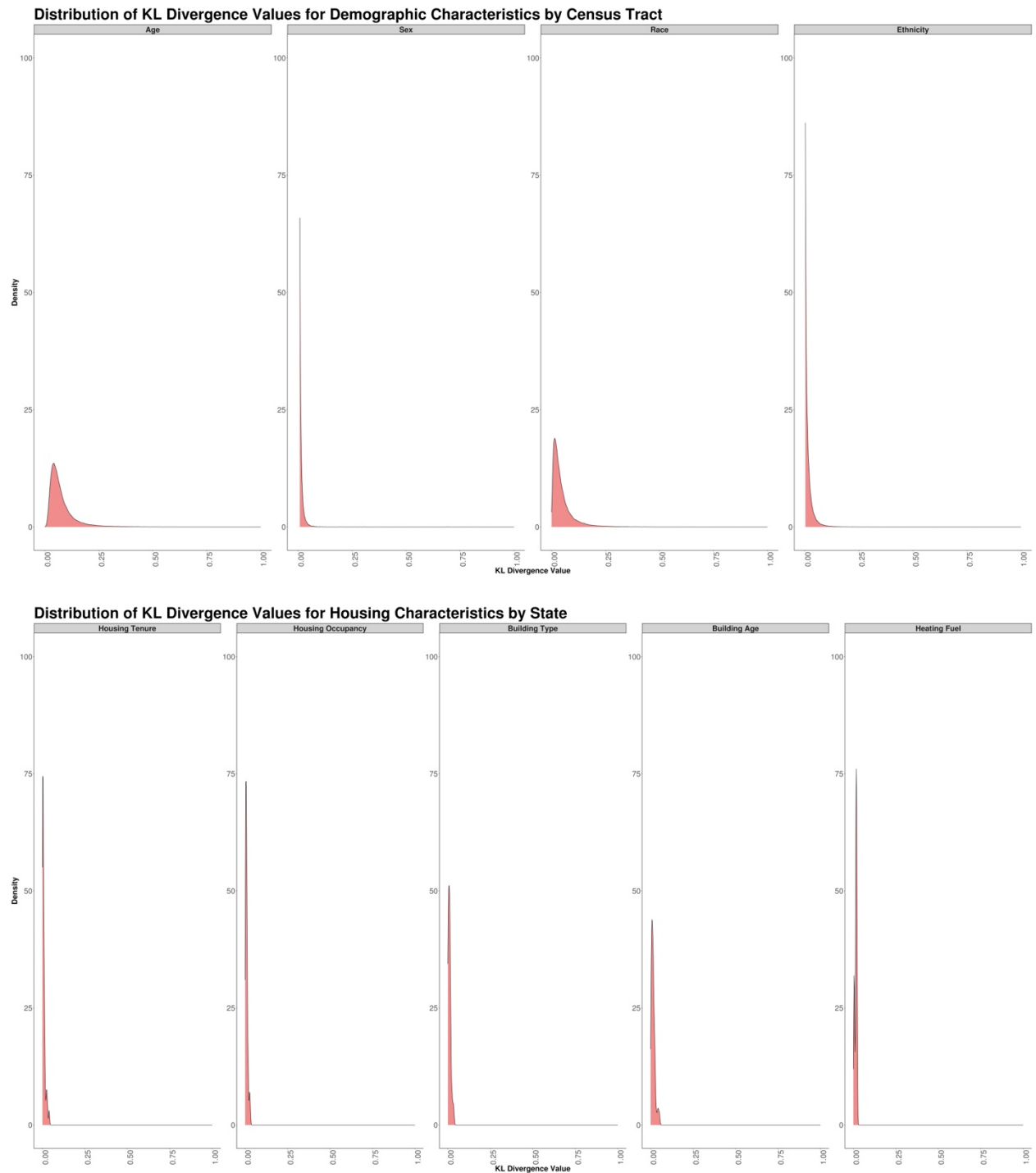


Figure 6. KL divergence statistic for (a) age, (b) sex, (c) race, and (d) ethnicity at the census tract level, and (e) housing tenure, (f) housing occupancy, (g) building type, (h) building age, and (i) heating fuel at the state level.

5. DISCUSSION.

This study developed the most up-to-date, comprehensive synthetic U.S. population designed to address complex public health challenges. By leveraging spatial microsimulation techniques, we generated approximately 330 million synthetic agents that statistically reflect the real population with respect to age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel at the census tract level. Unlike previous efforts that have been narrowly applied and/or inappropriately designed, our model stands out by capturing the intricate social, economic, infrastructural, and environmental interactions that underlie population health dynamics at a fine yet stable spatial resolution, ensuring its broader applicability. This broad applicability is further reinforced by our commitment to replicability, validation, and transparency, making it a reliable tool for public health researchers, policymakers, and practitioners aiming to better understand and address health and health inequities across the U.S.

Across the field of public health, spatial microsimulation has predominantly been applied to smoking,^{125–130} diabetes,^{131,132} obesity,^{131,133–135} and infectious disease.^{136,137} While some studies have expanded into areas like mental health,^{138–141} access to health-promoting resources,^{142–145} and environmental or climate health,^{146,147} the utility of these models remains limited. Models such as Edwards & Clarke’s SimObesity (2009)¹³³ and Tomintz et al.’s SimSalud (2013)¹⁴⁸ rely on regression and correlation analyses to identify linking variables specific to their areas of interest, such as age, gender, ethnicity, and educational attainment to predict obesogenic environments in Leeds, United Kingdom (U.K.) or smoking behaviors in Carinthia, Austria. The NREL’s LEAD tool,¹⁴⁹ which maps the spatial distribution of energy burdens across the U.S., relies on linking variables like household income, housing tenure, housing occupancy, building age, and heating fuel. However, in neglecting to use linking variables across a breadth of social, economic, infrastructural, and environmental characteristics, these models overlook the inherent connections between people and

the places they inhabit—smoking exposure is influenced by the age and the type of their homes,^{150,151} obesity is closely tied to the quality of their neighborhoods,^{152–154} and energy burdens disproportionately affect minoritized racial and ethnic groups even when controlling for household income.¹⁵⁵ This narrow focus not only fails to capture the complex interactions that shape population health dynamics, but also limits the applicability of these models beyond their specific contexts.

Very few models have been developed with the adaptability needed to address more diverse public health challenges at national scales. This shortfall is largely due to the computational demands required for spatial microsimulation at such large geographic scales. Models such as SimBritain in the UK¹⁵⁶ and SimAlba in Scotland¹⁵⁷ are notable for their thorough methodological approaches, but their models reflect the specific public health contexts of their respective countries, making direct comparisons to U.S.-based models difficult given the country’s size, diversity, and inequities. The most widely-used synthetic population in the U.S., SynthPop, employs IPF to downscale PUMS microdata to finer spatial resolutions using marginal distributions from the Census Bureau similarly to our approach.^{158,159} While this model has been applied to a broad range of public health issues—such as infectious disease,^{160–169} opioid abuse,^{170,171} obesity,¹⁷² and access to healthcare services,^{173–176} among others^{177–179}—significant limitations inherent to the baseline synthetic population itself raise concerns about its suitability for widespread public health research.

One of the fundamental issues with SynthPop is that it was constructed using methods suited for transportation modeling despite being developed for health modeling.^{158,159} SynthPop was built using the TRANSIMS software, which relies on linking variables optimized for transportation modeling, including the number of people in the household under the age of 18, household income, household size, household population, and vehicles available.^{158,159} However, these linking variables do not necessarily translate well to health modeling, missing critical predictors of health outcomes

such as age, sex, race, ethnicity, and housing conditions. When inappropriate linking variables are used in spatial microsimulation applied to health modeling, the resulting synthetic population may fail to capture the complex interactions that drive health outcomes, leading to biased models that oversimplify or misrepresent real-world population health dynamics.

SynthPop also uses census blocks as its geographic scale.^{158,159} While census blocks provide a finer spatial resolution than census tracts, they are relatively unstable over time due to population movement patterns, frequent boundary changes, and disclosure avoidance practices.^{110–113} This instability is particularly problematic if SynthPop continues to rely on data from 2000, 2010, or even 2015, as significant social, economic, infrastructural, and environmental shifts have occurred in recent years, which can further exacerbate inaccuracies in modeling when applied to contemporary public health issues.

Furthermore, the authors of SynthPop not only suggest that the resulting synthetic population reflects the real population with respect to its linking variables, but also assert that it “consists of a spatially and statistically accurate representation of each household and person in the U.S. with age, gender, income, race, educational attainment, and ethnicity”¹⁸⁰ despite that publicly available information does not show how the dataset performs relative to these variables at the block group level. The validation of TRANSIMS and SynthPop has been limited to just three variables—household income, household population, and household size—in just two counties using data from the year 2000.^{158,159} Given that specific variable and population attributes can significantly affect model performance, as supported by our findings, it is unreasonable to assume that these validation results hold true across other variables, geographic regions, spatial scales, or time periods without more extensive validation efforts.

These issues are compounded by a lack of transparency; there is no information available on the model inputs or parameters (e.g., data processing, initial weights, the number of iterations, and

the integerization process), but without this information, it is impossible to properly assess the model's assumptions or replicate its findings.^{158,159} As researchers have noted, “Simply describing the estimating method and reporting the inputs and outputs of the model are not good enough. Rigorous interrogation of the results must be attempted as to give the community of practice some confidence that the estimates are reliable” (Rose & Nagle, 2017).¹⁸¹ Despite these shortcomings, many studies continue to rely on SynthPop without interrogating its assumptions or considering how they might affect the validity of their findings.

5.1. Strengths.

The synthetic U.S. population presented here overcomes many of the limitations found in prior models in several ways. First, we specifically tailored our model for a wide range of public health applications by carefully selecting linking variables that capture the social, economic, infrastructural, and environmental characteristics that underlie population health dynamics. Using domain expertise, we identified key variables—age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel—that operate on multiple levels to shape individuals' lived experiences, including their exposures, their biological responses to those exposures, and their access to health-promoting resources. Second, because we selected the census tract level as our geographic scale, our model avoids issues related to the stability of small population sizes that are common when operating at the census block group level, while still capturing fine-grained spatial variations in health and health inequities that are missed when operating at the county level. This also has implications for both public health research and practice, where utilizing the census tract level facilitates integration with other major datasets and provides counties with actionable insights to target interventions at the community level. Third, we used the

most recent census data available, accounting for major social, economic, infrastructural, and environmental changes driven by urbanization, domestic and international migration, increasing racial and ethnic diversity, growing income inequalities, aging, and changing housing, environmental, and climatic conditions over the past 15-25 years. Fourth, we employed a widely-favored, efficient, and reliable modeling technique to ensure both the feasibility and replicability of the model for future applications. This approach, combined with extensive validation and transparency at every stage from data processing to model generation, gives future users confidence in the accuracy of the estimates while providing a clear understanding of the model's underlying assumptions, strengths, and potential limitations.

5.2. Limitations.

While addressing many limitations of previous approaches, the synthetic U.S. population presented here still faces certain challenges. One key issue is that, despite using “gold-standard” datasets and not identifying any patterns of missingness, drawing both the individual-level microdata and geographically-aggregated data from the U.S. Census Bureau may introduce unavoidable sampling biases that can permeate through the spatial microsimulation process. As Krieger (2019) highlights, data collected by the U.S. Census Bureau have historically been shaped by political and social forces, often excluding or misclassifying marginalized groups such as Indigenous populations or racialized minorities.^{182–184} Moreover, shifts in self-identification across census years can lead to inaccuracies that exacerbate the misrepresentation of these groups.¹⁸² These biases can carry through the spatial microsimulation process, affecting how well the synthetic U.S. population captures the diversity of the true population.^{185,186} While addressing these biases remains difficult without more

robust datasets, recent efforts by the U.S. Census Bureau have specifically targeted these groups to increase representativeness,¹⁸² potentially reducing its impact on our model.

Other key issues relate to the model parameters and modeling techniques chosen, which may account for the model's weaker performance when using linking variables with many subcategories, particularly in smaller, less diverse areas. Recent research shows that model parameters, such as the order of constraints in IPF and the integerization process, can impact model fit.¹²³ To mitigate this, we prioritized linking variables that were highly correlated with others early in the process while avoiding placing those with many subcategories at the end of the process, following the recommendations of Lovelace et al. (2015).¹²³ Additionally, although we used a pseudorandom number generator (TRS) to convert fractional weights into whole numbers during integerization, this approach can lead to minor discrepancies between the original and simulated data.⁹¹ Quasirandom number generators offer an alternative approach to reduce these discrepancies but struggle with incomplete data—something that IPF addresses through the use of a seed population.⁹² Lastly, while alternative methods to IPF such as combinatorial optimization may provide slight gains to model fit, they come with trade-offs such as reduced replicability and computational efficiency.⁸² Emerging techniques, including multilevel IPF, iterative proportional updating (IPU), penalized maximum entropy (P-MEDM), and fitness-based synthesis (FBS), offer promising refinements to IPF but require further validation before they can be applied reliably in large-scale public health research.⁵³ While our model performed extremely well across all validation metrics, future work should explore these emerging techniques to address the outlined issues with the model's performance.

6. CONCLUSIONS.

This study has demonstrated that our synthetic U.S. population is a powerful, adaptable tool for addressing complex public health challenges across the nation. By carefully selecting linking variables and geographic scales that reflect current social, economic, infrastructural, and environmental drivers of population health dynamics with geographic specificity, we built a model capable of delivering precise, localized insights. These features make it a valuable resource for public health researchers, policymakers, and practitioners aiming to improve health and health inequities in the U.S.

One simple truth lies at the core of any synthetic population: its potential can only be fully realized if it is built upon a solid foundation. Rigorous design, validation, and transparency are essential to ensuring that the model faithfully represents the real-world population. Without these, even the most “sophisticated” models risk generating incomplete or misleading results. We addressed this challenge by grounding our approach in the best practices of spatial microsimulation, ensuring that our synthetic U.S. population not only reflects the real population, but also serves as a reliable platform for further exploration and model-building.

Our synthetic U.S. population opens avenues of research that were previously inaccessible in three ways. First, it enables the study of joint distributions with geographic specificity. It moves beyond the flawed assumption that people who share single characteristics or reside in areas with similar marginal distributions of those characteristics experience health challenges in the same way. Instead, our model uniquely accounts for the complex, multi-dimensional, multi-level interactions that produce differential vulnerabilities at localized levels. Second, the flexibility of the model means that it can be readily integrated with other static or dynamic datasets or models, broadening its utility beyond what can be analyzed with the linking variables used in the original model-building process. Third, it can allow for the incorporation of target variables from the individual-level microdata or

from other external datasets—such as household relationships and family structures, language spoken and nativity status, housing, utility, and energy costs, occupation, industry, and employment details, insurance status and public assistance participation, and commuting patterns and transportation modes, time activity patterns—that were not part of the original model-building process but are otherwise difficult to attain. This facilitates the direct analysis of these target variables with geographic specificity, making it possible to design targeted interventions that address specific community needs. These capabilities can enable researchers and policymakers to not only understand vulnerabilities, but also test interventions that address these vulnerabilities in a virtual environment without the ethical or logistical constraints of real-world trials, across space and time. This tool therefore represents a significant step forward in advancing public health research and practice to address increasingly complex public health challenges in the U.S. with greater precision and care.

CHAPTER 3: LEVERAGING A SYNTHETIC POPULATION TO IDENTIFY CHILDREN AT HIGH RISK OF RESIDENTIAL LEAD EXPOSURE ACROSS THE UNITED STATES

Shivani Parikh^{a,b}, Lauren Ferguson^a, Brent A. Coull^a, John D. Spengler^a, Joseph G. Allen^a

^a Harvard T.H. Chan School of Public Health, 401 Park Drive, 4th Floor West, Boston, MA, 02215, USA

^b Harvard Graduate School of Arts and Sciences, 1350 Massachusetts Avenue, Cambridge, MA, 02138, USA

1. ABSTRACT.

The United States (U.S.) federal government has recently prioritized the development of residential lead exposure maps to identify areas and populations at highest risk of experiencing the toxic effects of lead-based paint exposure. Traditional approaches to generate these maps rely on overlapping quantiles of marginal distributions of risk proxies, including the percentage of young children and the percentage of older homes at the census block, census tract, or county level. However, these approaches are prone to the ecological fallacy—the assumption that relationships observed at the area level apply uniformly at the individual level—which can lead to misclassification of risk and misallocation of resources. To address these limitations, we developed a novel approach using a synthetic U.S. population to directly model the joint distribution of young children living in older homes at the census tract level. This novel joint distribution approach identified 73% more at-risk census tracts—where an estimated 3.28 million additional at-risk children reside—than the traditional marginal distribution approach. Importantly, our spatial autoregressive (SAR) models showed that these undetected areas were not random but systematically associated with census tracts containing higher proportions of uninsured, Native American, and White populations, which were respectively 73%, 18%, and 17% more likely to be undetected by traditional approaches. These findings reveal critical inequities in the ability of traditional approaches to capture at-risk populations, particularly in low-density areas. Further, our generalized linear mixed-effects (GLMM) models uncovered that uninsured and low-income children were respectively 62% and 11% more likely to live in older homes compared to their insured and high-income counterparts. White children were also more likely to reside in older homes compared to children from all other racialized groups, in direct contrast to existing literature which has consistently shown that minoritized populations experience disproportionately higher blood lead levels. This discrepancy underscores the inadequacy of housing age as a sole risk proxy without accounting for disparities in

housing quality across racialized groups. By leveraging the spatially explicit joint distributions inherent in our synthetic U.S. population, we were able to uncover not only where there are high concentrations of young children living in older homes, but also which subpopulations of young children are more likely to reside in older homes. This nuanced understanding demonstrates that our synthetic U.S. population can provide a more equitable and precise framework for identifying residential lead exposure hotspots and enable the development and deployment of targeted public health interventions that protect the most vulnerable populations across the U.S.

2. INTRODUCTION.

Lead exposure continues to be one of the most pervasive public health challenges in the United States (U.S.). Throughout the 20th century, lead was introduced as a key component of products including paint, plumbing, and gasoline because of its durability and versatility.¹⁸⁷ However, decades of research has shown that lead is not only highly persistent in the environment but also extremely dangerous to human health, acting as a potent neurotoxin. Despite regulatory bans enacted on lead-based paint in 1978, lead-based pipes in 1986, and leaded gasoline in the 1990s, these regulations did not require the immediate removal of lead-based materials from existing infrastructures, meaning that the burden of remediation largely fell on homeowners, building owners, and municipalities—many of whom lacked the resources needed to intervene.¹⁸⁸ Therefore, lead remains embedded in the dust, soil, water, and air to which we are exposed every day, continuing to pose a significant threat to population health.¹⁸⁹

Children are particularly vulnerable to the harmful effects of lead exposure.¹⁹⁰ Physiologically, children under the age of six absorb lead more efficiently than adults, and their developing nervous systems are more susceptible to its toxic effects.¹⁹⁰ Even low levels of lead

exposure in early childhood have been associated with cognitive deficits, developmental delays, and behavioral problems that can severely impair long-term health and future educational and economic opportunities.^{190–193} These challenges come at a significant cost—not only to children and their families, but also to society—with billions of dollars lost annually to increased healthcare utilization, special education needs, and reduced workforce productivity.^{194–196}

Homes built before regulatory bans on lead-based products went into effect are a major source of lead exposure for children under the age of six, and therefore, a major target for public health interventions. Over time, lead-based paint deteriorates, releasing toxic dust and paint chips that accumulate on surfaces such as floors, windowsills, and furniture. Everyday activities like cleaning, opening and closing windows, and playing on the floor can disturb these particles, increasing the likelihood of inhalation or ingestion by children who frequently engage in hand-to-mouth behaviors and come in close contact with the floor. The soil around these homes is often contaminated with lead from deteriorating exterior paint or residual leaded gasoline, creating further risks for children playing outdoors. Additionally, many of these homes, often found in older cities and towns, are serviced by lead-based pipes that can leach lead into drinking water as they corrode. An estimated 34 million homes still contain lead-based paint and 9 million service lines with lead-based pipes remain in use, highlighting the persistent and multifaceted risks for young children living in older homes, particularly those who belong to low-income and minoritized communities^{197–202} where older, substandard infrastructures are more prevalent^{29,30,34–40} and resources for remediation are often limited.²⁰³

Blood lead level (BLL) testing is considered the “gold standard” method for quantifying children’s total lead exposures from all sources, measuring actual exposures rather than merely estimating potential ones.²⁰⁴ This method also offers a standardized way to monitor individual- and population-level responses to large-scale interventions over time.²⁰⁴ Despite these benefits, however,

BLL testing is subject to several limitations that impede its effectiveness in capturing the full scope of lead exposure and targeting public health interventions across the U.S.:

- First, only 11 states and the District of Columbia mandate universal BLL testing for children, largely due to the high monetary and administrative costs of drawing and analyzing blood. As a result, most states rely on targeted testing in areas deemed “high-risk” (which hinges on the assumption that those areas are correctly classified as “high-risk” in the first place), or only provide recommendations for testing.^{205–209} In many cases, this leads to gaps in testing or no testing at all, leaving significant portions of the child population unmonitored.
- Second, even when testing does occur, data collection is often incomplete or unrepresentative. Poor and minoritized communities, despite facing disproportionately higher risks of lead exposure, are frequently under-sampled due to limited access to healthcare services, lack of awareness, and insufficient outreach efforts.^{210–212} These sampling biases result in data that fail to capture the full scope of lead exposure across populations.
- Third, due to data privacy regulations, states often aggregate BLL data at the county level, which can obscure important individual or spatial variations that may be present within counties. While some states aggregate data at the census tract level, the small number of individuals tested—common in many areas—necessitates data suppression for those census tracts. It is even rarer for states to make individual-level BLL data available, and when it is, key individual-level characteristics such as race, ethnicity, household income, and housing conditions are typically masked to adhere to data privacy regulations.

- Fourth, data sharing practices are inconsistent across states, as each sets its own reporting regulations, protocols, and standards. This fragmented approach results in a patchwork of data, making it difficult to get a comprehensive national picture of exposure risks.
- Fifth, BLL data alone does not provide information about the specific sources of lead exposure, making it difficult to target interventions that would protect against those specific sources of lead exposure.

The Federal Lead Task Force—which includes members from the U.S. Environmental Protection Agency (EPA) Department of Health and Human Services (HHS), Agency for Toxic Substances and Disease Registry (ATSDR), Centers for Disease Control and Prevention (CDC), Food and Drug Administration (FDA), Department of Housing and Urban Development (HUD), Occupational Safety and Health Administration (OSHA), and the U.S. Department of Agriculture (USDA)—launched the Federal Action Plan to Reduce Childhood Lead Exposure in 2018 in recognition of these issues.²¹³ **A core focus of this action plan is on the development of residential lead exposure maps as tools to enhance BLL surveillance, inform primary prevention, and target areas for lead mitigation, enforcement, education, and outreach programs.**^{214–216} Accurate mapping is particularly crucial in light of efforts by HUD to direct lead hazard reduction grants.²¹⁵ These grants are specifically allocated to state and local governments to identify and control lead-based paint (LBP) hazards in privately owned pre-1978 homes occupied by low-income families.²¹⁵ However, the funding available is insufficient to address the risks across all homes fitting this profile; a 1000-fold increase would be needed to cover all privately owned pre-1978 homes occupied by low-income families.²¹⁵ **To this end, HUD encourages grantees to prioritize areas with high concentrations of pre-1978 housing, young children, and low-income families as per Executive Order 14008’s Environmental Justice Initiative.** This underscores the importance of accurate residential lead exposure mapping, as effective allocation of

limited resources to high-priority areas depends on the accuracy of these maps.²¹⁵ Consequently, as stressed in a joint statement by members of the Federal Lead Task Force, “as public health practitioners rely increasingly on lead maps to inform risk identification and management, it is vital that the underlying science is sound and transparent.”²¹⁶

In support of these federal initiatives, public health researchers traditionally rely on proxies for residential lead-based paint exposure risk. The most widely-used source of data for these proxies is the American Community Survey (ACS), a nationally representative survey conducted annually by the U.S. Census Bureau. To protect respondent privacy, the ACS reports data as marginal distributions, which represent aggregates of single characteristics at various geographic levels. Because these variables are measured on different scales, quantile classification is used to standardize them, dividing the data into equal parts to ensure that each variable contributes equally to the analysis.^{73,217–219} These variables are then overlaid on a map to pinpoint areas where the top quantiles overlap with one another.^{73,217–219} In this context, researchers may assess where the top quantile for the percentage of the population living in homes built before 1980 overlaps with the top quantiles for the percentage of the population under the age of five, the percentage of the population in poverty, the percentage of the population identifying as a racial or ethnic minority, and/or the percentage of the population renting their homes at the census block, census tract, or county levels in order to identify residential lead exposure hotspots, recognizing that these characteristics often compound risks associated with lead exposure in homes.^{73,217–219}

However, these traditional approaches assume that the mere presence of these characteristics indicates that they intersect with one another in meaningful ways, which is not often the case.

Specifically, while marginal distributions capture how each characteristic occurs independently of all other characteristics within an area, they do not reflect joint distributions that capture how those characteristics co-occur together within that same area.

For example, just because an area has a high concentration of young children and a high concentration of older homes (marginal distributions) does not mean that the area has a high concentration of young children who *live in* older homes (joint distribution). The failure to distinguish between marginal and joint distributions is known as the ecological fallacy, where relationships observed at the group level are mistakenly assumed to apply at the individual level, even though these relationships may differ significantly in magnitude and/or direction.^{63,68,69} Exacerbating this issue is the need to use quantile classification when working with marginal distributions as described above. **Dividing the data into these broad categories—such as tertiles, quartiles, or quintiles—can obscure subtle but important variations in risk across areas.**^{73,217–219} The limitations of traditional approaches to residential lead exposure mapping may lead to the misidentification of residential lead exposure hotspots and, consequently, the misallocation of interventions and resources for addressing lead-based paint hazards.

To address the limitations of these traditional approaches, we utilize a novel synthetic U.S. population we previously generated that uniquely models joint distributions of multiple characteristics across approximately 330 million synthetic agents at the census tract level (Parikh et al., 2024), to produce a new map of residential lead exposure hotspots across the U.S.^{53–55} In this study, we specifically leverage this synthetic U.S. population to: (1) identify where young children living in older homes are concentrated, (2) determine whether traditional approaches for residential lead exposure assessment may misclassify young children living in older homes, (3) determine whether traditional approaches for lead exposure assessment may systematically misclassify young children living in older homes based on the specific demographic characteristics of the areas where they are concentrated, and (4) analyze whether specific demographic characteristics increase the likelihood of young children living in older homes. Through this analysis, we aimed to demonstrate that our synthetic U.S. population can offer a novel approach to more effectively pinpoint

communities that may face compounding risks from lead exposures within their homes. The broader implication of this novel approach lies in its capacity to more effectively guide the development and deployment of interventions and resources for addressing lead-based paint hazards to the communities that need them most.

3. METHODS.

3.1. Data sources.

This study relies on two primary datasets: the traditional American Community Survey (ACS) dataset, which provides marginal distributions of single characteristics across the U.S. population at the census tract level, and a novel synthetic U.S. population dataset, which provides joint distributions of multiple characteristics across the U.S. population at the census tract level.

3.1.1. ACS dataset.

For the traditional ACS dataset, we extracted data on age, sex, race, and ethnicity from Table S0601; household income from Table S1901; housing occupancy from Table S2501; and housing tenure, building type, building age, and heating fuel from Table DP04, all at the census tract level for the U.S. and Washington D.C. These data were processed by converting percentages to counts, harmonizing household-level and individual-level variables, and addressing discrepancies through proportional adjustments to maintain consistency across variable subcategories. The methods used to process the traditional ACS dataset are described in detail elsewhere (Parikh et al., 2024).

3.1.2. Synthetic U.S. population dataset

The methods used to construct our novel synthetic U.S. population dataset are similarly described in detail elsewhere (Parikh et al., 2024). Briefly, we employed a static spatial microsimulation approach using Iterative Proportional Fitting (IPF), a deterministic reweighting method known for its efficiency and reliability. This involved constraining individual-level microdata from the 2022 1-Year ACS Public Use Microdata Sample (PUMS) to their known marginal totals from the 2022 5-Year American Community Survey (ACS) across age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel at the census tract level. To initialize the IPF process, individuals were assigned equal weights, which were iteratively adjusted up to 50 iterations. Once convergence was reached, where further adjustments aligning the microdata with their known marginal totals resulted in negligible changes defined as a maximum tolerance threshold of $1.5e-8$ or as the best available adjustments at 50 iterations, the fractional weights were integerized using the Truncate, Replicate, Sample (TRS) method to produce integer weights.⁹¹ Each individual was then replicated according to their corresponding integer weights. The resulting synthetic U.S. population dataset underwent extensive internal and external validation to ensure its accuracy and applicability for further public health analyses as described in Parikh et al. (2024).

3.2. Statistical analysis.

3.2.1. Comparing census-tract-level risk classification.

To assess potential exposure misclassification by the traditional marginal distribution approach compared to our novel joint distribution approach, we classified risk at the census tract level under each approach using quartile-based thresholds.

To model a simple version of the traditional marginal distribution approach, we calculated separate quartiles of the percentage for the population under the age of five and quartiles of the percentage of homes built before 1980 using their respective marginal distributions from the ACS dataset, and then overlaid these quartiles. Census tracts were classified as “high-risk” for lead exposure if both marginal distributions of the percentage of the population under the age of five and the percentage of homes built before 1980 were in the fourth quartile.

To enable the direct comparison of this traditional marginal distribution approach against our novel joint distribution approach, we computed quartiles of the percentage of children under the age of five *living in* homes built before 1980 using this joint distribution within the synthetic U.S. population dataset, where census tracts in the fourth quartile were classified as “high-risk” for lead exposure.

3.2.1.1. Quantifying the extent of exposure misclassification.

We assessed the extent of exposure misclassification using three different metrics: (1) the intersection over union (IoU), which quantifies the overlap between census tracts classified as “high-risk” by the traditional marginal distribution approach and our novel joint distribution approach; (2) the percentage of census tracts identified as “high-risk” under the traditional marginal distribution approach but not by our novel joint distribution approach; and (3) the percentage of tracts identified as “high-risk” by our novel joint distribution approach but not by the traditional marginal distribution approach. We also estimated the number of children under the age of five living in pre-1980 homes who resided in census tracts that were identified as “high-risk” by our novel joint distribution approach but not by the traditional marginal distribution approach. To evaluate the stability of our findings, we conducted a sensitivity analysis by relaxing the assumptions of what

constitutes “high-risk” under the traditional marginal distribution approach. Whereas we originally defined “high-risk” as if *both* marginal distributions were in the fourth quartiles, our first sensitivity analysis defined “high-risk” as if *either* marginal distribution was in at least the third quartile and the second scenario defined “high-risk” as if *both* marginal distributions were in at least the third quartile, as detailed in **Appendix A**.

3.2.1.2. Quantifying the nature of exposure misclassification

We further explored the nature of exposure misclassification, or in other words, whether children under the age of five living in homes built before 1980 were systematically more likely to go undetected as “high-risk” when using the traditional marginal distribution approach compared to our novel joint distribution approach based on specific population characteristics of the tracts in which they resided.

To test this, we first aggregated our full synthetic U.S. population dataset to the census tract level, ensuring that all variables operated at the same geographic scale. Next, we assessed the presence, degree, and direction of spatial autocorrelation in the undetected tracts using Moran’s I test. This step was critical because spatial autocorrelation—defined as when similar values cluster geographically rather than being randomly distributed—violates the independence assumption required for unbiased estimates in standard logistic regression models. The Moran’s I statistic is calculated as follows:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

Where:

- I is Moran’s I statistic;
- N is the number of census tracts;

- W is the sum of all spatial weights w_{ij} ;
- x_i and x_j are the values of the variable in tracts i and j ;
- $\underline{x}$ is the mean value of the variable;
- w_{ij} is the spatial weight between tracts i and j , defined using queen's contiguity, where $w_{ij} = 1$ if tracts i and j share a border or a vertex and $w_{ij} = 0$ otherwise.

Given that the results of the Moran's I test indicated spatial autocorrelation of undetected tracts (Moran's I = 0.109, $p < 0.001$), we employed spatial autoregressive (SAR) models, which accounts for the influence of neighboring tracts through a spatial lag term. These models assessed the likelihood of tracts being undetected by the traditional approach compared to our novel approach as a function of the tract's race, ethnicity, household income, housing tenure, building type, and health insurance status distributions—characteristics selected a-priori. To address potential multicollinearity, we specified two models: one including non-White racialized groups and all other predictors, and another including the White racialized group and all other predictors. The SAR model is expressed as:

$$\text{logit}(P(Y_j = 1)) = \rho WY_j + X_j\beta$$

Where:

- Y_j is the binary outcome for census tract j , representing the log-likelihood of not detecting a census tract at “high-risk” of residential lead-based paint exposure using the traditional marginal distribution approach compared to our novel joint distribution approach;
- ρ is the spatial autoregressive parameter, capturing the influence of neighboring tracts;

- W is the spatial weight matrix that defines the neighboring structure between tracts, where we used a K-nearest neighbors approach with $k = 6$, and specified that the centroid of each census tract determined its neighbors;
- WY is the spatial lag of the dependent variable, representing the weighted average of neighboring tracts' outcomes;
- X_j is matrix of predictor variables for census tract j , including the tract's race, ethnicity, household income, housing tenure, building type, and health insurance status distributions
- β represents the coefficients for the predictor variables.

3.2.2. Identifying individual-level risks of young children living in older homes.

We then sought to use the synthetic U.S. population dataset to identify whether specific characteristics of children under the age of six (an important distinction from our previous analyses considering children under five, which were restricted by the ACS dataset's inability to modify specific age categories, despite the CDC's identification of 0-6 as an age group particularly vulnerable to lead exposure) were associated with their likelihood of living in homes built before 1980. Using the flexibility of our synthetic U.S. population dataset, we examined how multiple characteristics (poverty status, health insurance status, race, ethnicity, and housing conditions) interact within this age group. This analysis was not possible with the ACS dataset, which lacks the ability to explore the co-occurrence of multiple characteristics within a given age group.

To model this, we filtered the synthetic U.S. population dataset for children under the age of six and applied a generalized linear mixed effects model (GLMM), with living in a home built before 1980 as the outcome. This model is well-suited for accounting for both individual-level characteristics, and the hierarchical nesting of individuals within census tracts and census tracts

within states. We included fixed effects for poverty status, health insurance status, race, ethnicity, housing tenure, and building type—characteristics selected a-priori for their potential impact on lead exposure risks. We also included nested random effects to capture the unobserved heterogeneity at the census tract and state levels by allowing each geographic unit to have its own intercept, thus adjusting for shared variability among individuals in the same area and accounting for localized differences in lead exposure risks. To explore how poverty might modify these relationships, we also stratified the analysis by poverty status, defining our threshold at 100% the federal poverty level (FPL) or an annual household income of \$25,000. The model was fit using the following formula:

$$\text{logit}\left(P(Y_{ijk} = 1)\right) = X_{ijk}\beta + u_j + v_j + w_{ijk}$$

Where:

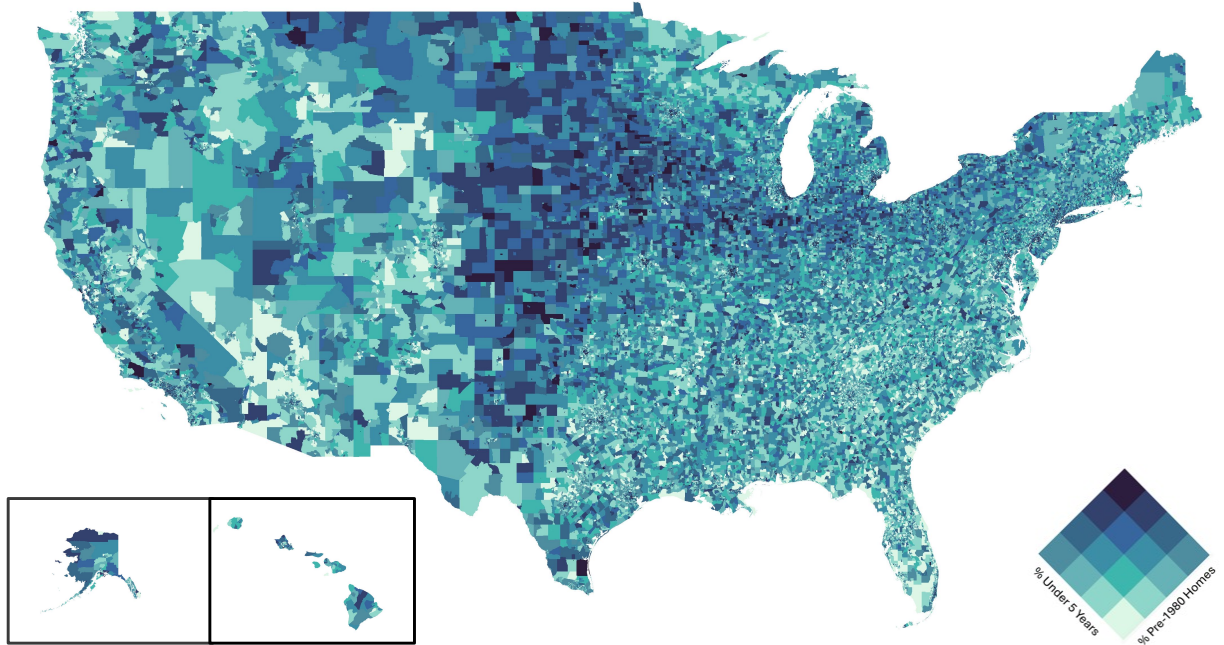
- Y_{ijk} is the binary outcome for individual i in census tract j within state k , representing the log-likelihood of a child under the age of six living in a home built before 1980;
- X_{ijk} is the matrix of fixed-effect predictor variables for individual i in census tract j within state k , including poverty status, health insurance status, race, ethnicity, housing tenure, and building type;
- β represents the coefficients for the fixed-effect predictor variables;
- u_k is the random effect associated with state k , capturing unobserved variability between states;
- v_j is the random effect associated with census tract j , capturing unobserved variability between census tracts;
- w_{ijk} is the random effect associated with individuals within census tracts, capturing unobserved heterogeneity at the individual level.

4. RESULTS.

4.1. Comparing census-tract-level risk classification.

We found that areas with high concentrations of young children and high concentrations of older homes identified by the traditional marginal distribution approach (**Figure 7a**) differed fundamentally from areas with high concentrations of young children living in older homes identified by our novel joint distribution approach (**Figure 7b**).

Traditional Marginal Distribution Map for Residential Lead Exposure Risk



Novel Joint Distribution Map for Residential Lead Exposure Risk

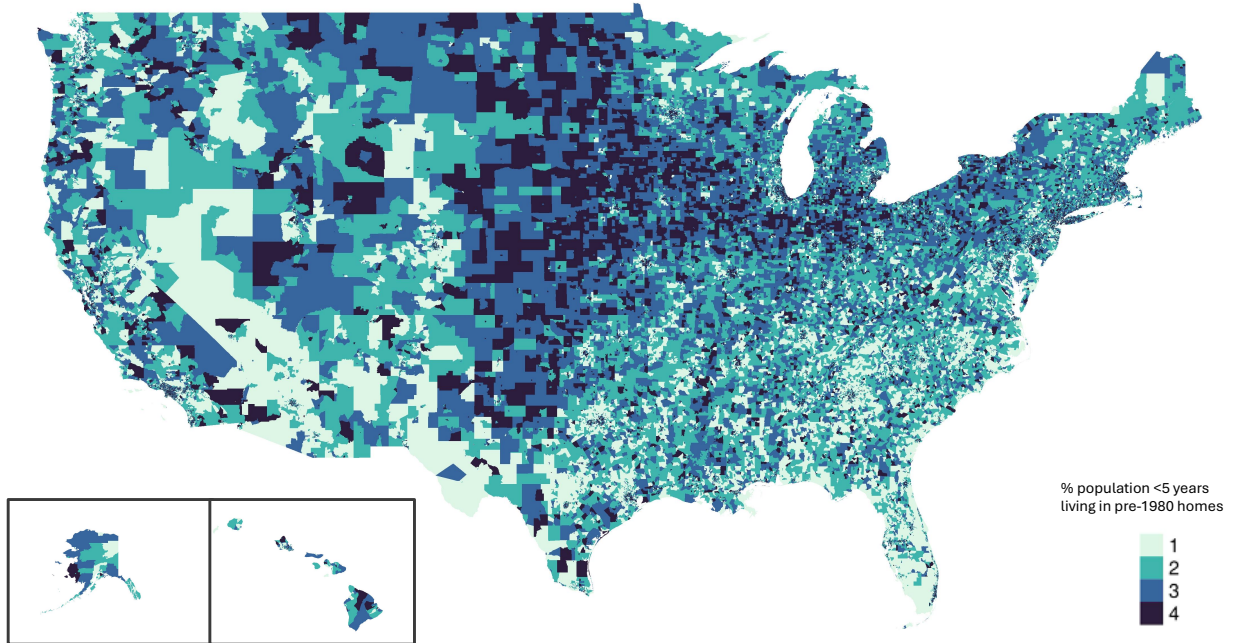


Figure 7. Visual comparison of census tracts identified as “high-risk” for lead exposure across the U.S. by (a) the traditional marginal distribution approach using the ACS dataset versus (b) our novel joint distribution approach using the synthetic U.S. population dataset.

4.1.1. Quantifying the extent of exposure misclassification.

When we quantified the extent of exposure misclassification, we first observed an IoU value of 26.8%, suggesting a significant divergence between the two approaches in identifying “high-risk” areas. We then examined how many tracts were classified as “high-risk” by one approach but not the other. Only 1.4% of tracts identified as “high-risk” by the traditional marginal distribution approach were not identified as “high-risk” by our novel joint distribution approach; conversely, 73.2% of tracts identified as “high-risk” by our novel joint distribution approach were not identified as such by the traditional marginal distribution approach. Assuming that the synthetic U.S. population provides an accurate quantification of the joint distribution of young children living in older homes at the census tract level, this difference translates to an estimated 3.28 million children under the age of five living in homes built before 1980 who reside in tracts that would have been overlooked by the traditional marginal distribution approach, representing a significant expansion in the identification of “high-risk” areas using our joint distribution approach. These differences persisted when we relaxed the criteria for defining “high-risk” in the traditional approach compared to our novel approach (**Tables 1A-2A in Appendix A**).

Importantly, however, the extent of these differences varied by state. Traditional marginal distribution approaches struggled to detect young children living in older homes in both rural states such as Nebraska and Iowa and in urban states with diverse populations such as New York, Illinois, Massachusetts, and California. In contrast, these traditional approaches struggled less in more homogenous states, such as Vermont, Nevada, Maine, and Alaska, as well as many Southern states (**Figure 8**).

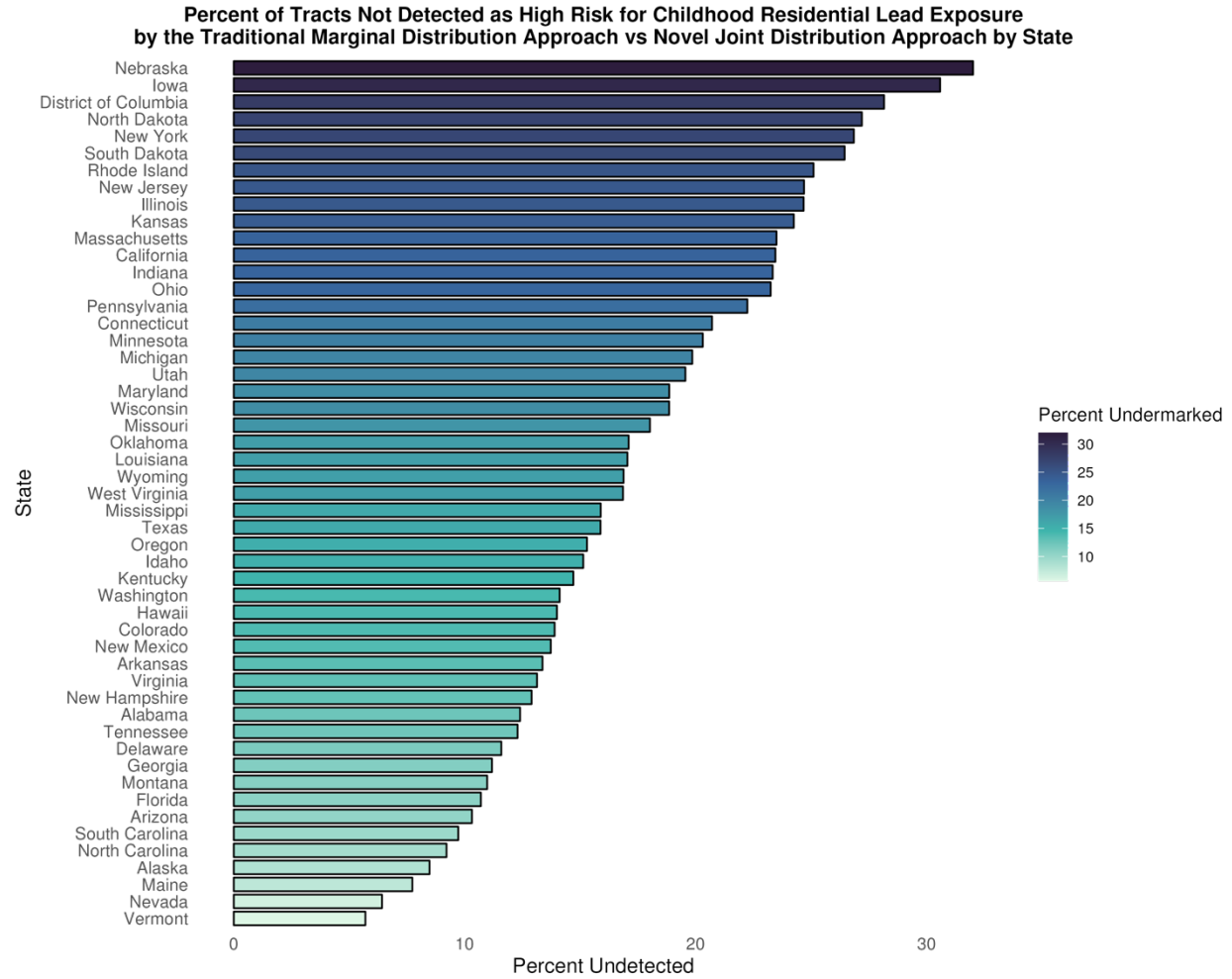


Figure 8. Percent of tracts not detected as “high-risk” for childhood residential lead exposure by the traditional marginal distribution approach compared to our novel joint distribution ordered by state.

4.1.2. Quantifying the nature of exposure misclassification.

We then found that children under the age of five living in homes built before 1980 were systematically more likely to go undetected as “high-risk” when using the traditional marginal distribution approach compared to our novel joint distribution approach based on specific population characteristics of the tracts in which they resided. The SAR model results revealed that tracts with higher proportions of White, Native American, and uninsured populations were

significantly more likely to go undetected as “high-risk”. Specifically, for every 1% increase in the proportion of White, Native American, and uninsured populations each in a given census tract, the odds of that tract going undetected as “high-risk” increased by 17% (OR: 1.174 [1.139, 1.211]), 18% (OR: 1.179 [1.009, 1.379]), and 73% (OR: [1.529, 1.953]) respectively, controlling for spatial dependence and other variables in the model (**Table 8**). These patterns also persisted in sensitivity analyses (**Tables 3A-4A in Appendix A**).

Table 8. Results of the SAR model presented for the full population.

Variable	Non-White	White
Intercept	2.889**** [2.762, 3.023]	2.508**** [2.379, 2.643]
% <100% FPL	0.818**** [0.765, 0.874]	0.795**** [0.746, 0.847]
% Uninsured	1.728**** [1.529, 1.953]	1.725**** [1.528, 1.946]
% White	—	1.174**** [1.139, 1.211]
% Black	0.844**** [0.816, 0.873]	—
% Asian	0.930* [0.860, 1.006]	—
% Native	1.179** [1.009, 1.379]	—
% Hispanic	0.834**** [0.806, 0.862]	0.918**** [0.888, 0.949]
% Rent	0.798**** [0.757, 0.841]	0.806**** [0.765, 0.850]
% SFH	0.783**** [0.751, 0.817]	0.778**** [0.747, 0.811]

4.2. Identifying individual-level risks of young children living in older homes.

Table 9 presents the descriptive statistics of the 22.47 million children under the age of six in the synthetic U.S. population. The average household income for families with children under the

age of six is \$112,620, with 86% of children living in households above the FPL and 14% of children living in households below it. Health insurance coverage varies, with 53% of children being privately insured, 43% publicly insured, and 4% without any insurance. Among these children, 58% are White, 13% are Black, 6% are Asian, 1% are Native American, and 22% fall into other racial categories (which includes those identifying as a mixed or other race). The majority are non-Hispanic (73%) while the remainder are Hispanic (27%). Housing characteristics reveal that a majority (62%) of children live in owner-occupied homes, and most (72%) reside in single-family homes. Notably, there is an even split between children living in homes built before 1980 (49%) and those in homes built after 1980 (51%).

Table 9. Descriptive statistics of children under the age of six in the synthetic U.S. population dataset.

Variable	N = 22,470,473
Federal Poverty Level	
>100% FPL (>\$25,000)	19,432,530 (86%)
<100% FPL (<\$25,000)	3,037,943 (14%)
Health Insurance Coverage	
Privately insured	11,972,409 (53%)
Publicly insured	9,595,418 (43%)
Uninsured	902,646 (4.0%)
Race	
White	13,049,068 (58%)
Asian	1,241,767 (6%)
Black	2,895,051 (13%)
Native	246,252 (1%)
Other	5,038,335 (22%)
Ethnicity	
Non-Hispanic	16,394,835 (73%)
Hispanic	6,075,638 (27%)
Housing Tenure	

Own	13,864,663 (62%)
Rent	8,605,810 (38%)
Building Type	
Single family home	16,256,952 (72%)
Apartment	5,024,890 (22%)
Mobile Home	1,188,631 (5.3%)
Building Age	
Pre-1980	11,015,416 (49%)
Post-1980	11,455,057 (51%)

We used this analytic sample to evaluate how social, economic, infrastructural, and environmental characteristics influenced the likelihood of children under six living in homes built before 1980 (**Table 10**). Income emerged as playing an important role in this relationship. Children from households earning below 100% of the FPL were 11% more likely to live in older homes compared to those from households earning above 100% the FPL (OR = 1.105 [1.101, 1.109]). Health insurance status also played an important role in determining which children were more likely to live in older homes, a relationship which varied with household income level. Children who were uninsured (OR = 1.619 [1.616, 1.621]) or publicly insured (OR = 1.330 [1.324, 1.336]) were significantly more likely to live in homes built before 1980 compared to those with private insurance. The effect of insurance status was more pronounced in lower-income households, where children who were uninsured (OR = 1.871 [1.856, 1.889]) or publicly insured (OR = 1.656 [1.632, 1.680]) had even higher odds of living in older homes compared to their privately insured counterparts.

Racial and ethnic differences in the likelihood of children under the age of six living in homes built before 1980 were also evident. Compared to White children, Asian children were 50% less likely (OR = 0.497 [0.491, 0.503]), Black children were 21% less likely (OR = 0.789 [0.785, 0.794]), and Native American children were 6% less likely (OR = 0.936 [0.925, 0.947]) to live in older homes. However, these relationships showed more nuanced patterns when stratified by

household income level. For young Asian and Black children in lower-income households, the likelihood of living in older homes remained lower compared to their White counterparts, though the disparity was reduced (OR = 0.748 [0.717, 0.779]; OR = 0.998 [0.984, 1.011]). Interestingly, young Native American children in lower-income households were even less likely to live in older homes than their White counterparts (OR = 0.516 [0.478, 0.555]), but those in higher-income households were more likely to live in older homes (OR = 1.054 [1.042, 1.067]).

Finally, housing differences were also evident, where children in rented homes were more likely to reside in older homes compared to those in owned homes (OR = 1.166 [1.163, 1.169]), and children in apartments (OR = 0.433 [0.429, 0.437]) or mobile homes (OR = 0.193 [0.187, 0.199]) were less likely to reside in older homes compared to those in single-family homes.

Table 10. Results of the GLMM presented for the full population and stratified by poverty level.

Variable	Full	<100% FPL	>100% FPL
Intercept	1.090** [1.039, 1.142]	1.827**** [1.682, 1.973]	1.084**** [1.047, 1.121]
<100% FPL	1.105**** [1.101, 1.109]	—	—
Uninsured	1.619**** [1.616, 1.621]	1.871**** [1.856, 1.889]	1.638**** [1.635, 1.641]
Public insurance	1.330**** [1.324, 1.336]	1.656**** [1.632, 1.680]	1.342**** [1.335, 1.348]
Asian	0.497**** [0.491, 0.503]	0.748**** [0.717, 0.779]	0.465**** [0.458, 0.471]
Black	0.789**** [0.785, 0.794]	0.998 [0.984, 1.011]	0.772**** [0.767, 0.778]
Native	0.936**** [0.925, 0.947]	0.516**** [0.478, 0.555]	1.054**** [1.042, 1.067]
Other	1.033**** [1.029, 1.036]	1.279**** [1.266, 1.292]	1.011**** [1.007, 1.0149]
Hispanic	1.057**** [1.053, 1.061]	0.976*** [0.962, 0.990]	1.063**** [1.059, 1.067]
Rent	1.166**** [1.163, 1.169]	1.014**** [1.003, 1.025]	1.199**** [1.195, 1.202]

Apartment	0.433**** [0.429, 0.437]	0.365* [0.354, 0.376]	0.420**** [0.415, 0.424]
Mobile Home	0.193**** [0.187, 0.199]	0.101**** [0.084, 0.118]	0.187**** [0.181, 0.194]

5. DISCUSSION.

This study aimed to demonstrate that our novel synthetic U.S. population—in providing access to joint distributions of key risk proxies for residential lead-based paint exposures with geographic precision—can be used as a powerful tool for identifying both who is at risk and where those most at risk live. Our census-tract-level analysis revealed that 73% of areas identified as “high-risk” in our novel joint distribution approach (those with high concentrations of individuals under the age of five living in homes built before 1980) were not identified as such by the traditional marginal distribution approach (those with high concentrations of individuals under the age of five and high concentrations of individuals living in homes built before 1980). This discrepancy translates to up to 3.28 million young children who may go undetected in risk assessment and mitigation efforts when traditional methods are used. Moreover, our findings suggest that tracts with high concentrations of White, Native American, and uninsured individuals were systematically more likely to go undetected by traditional approaches. Further individual-level analyses revealed that children under the age of six from households earning below the FPL were 11% more likely to live in older homes, and uninsured children were 62% more likely to reside in older homes compared to those with private insurance, an effect which was more pronounced for children from households earning below the FPL. These findings highlight the ability of our synthetic U.S. population to uncover hidden, compounding risks of residential lead-based paint exposures and better inform the development and deployment of interventions and resources for addressing those hazards to the communities that need them most.

Our analysis shows that by utilizing our novel joint distribution approach, we can expand the ability to identify census tracts at risk of residential lead-based paint exposure by 73%, potentially uncovering an additional 3.28 million children who may be at risk compared to traditional marginal distribution approaches. These traditional approaches—including those employed by the CDC’s Lead Exposure Risk Index (LERI),²¹⁷ EPA’s Pilot Study of Targeting Elevated Blood-Lead Levels in children,²²⁰ EPA’s EJSCREEN Lead Paint Environmental Justice Index,⁷³ Vox’s Lead Risk Score,²¹⁸ and the Schultz model²¹⁹—mask the complex interactions that may compound risks in specific subpopulations by focusing on isolated characteristics such as the percentage of the population under the age of five, the percentage of the population in poverty, the percentage of the population identifying as a racial or ethnic minority, and/or the percentage of the population renting their homes.²¹⁶

Recent work by Xue et al. (2022)²²¹ and Stanek et al. (2024)²²² has advanced these methods by leveraging actual BLL data in geospatial analyses in Michigan and Ohio. By using two main methods—the top 20th percentile method, which flags census tracts with the highest proportions of elevated BLLs, and the Getis-Ord Gi Cluster Analysis*, which detects spatial clusters based on the surrounding distribution of elevated BLLs—they provide a more precise identification of risk by focusing on direct measures rather than proxy measures in these two states. When Zartarian et al. (2024) compared total lead exposure hotspots identified using these methods to the residential lead exposure hotspots identified using traditional marginal distribution approaches at the national scale, the authors unsurprisingly found only moderate agreement (Cohen’s kappa 0.49–0.63), with the highest agreement observed between these methods and a new EPA-ORD random forest model that uses ACS data as inputs (yet to be published).²²³ Machine learning models similar to this random forest model have gained traction in recent years due to their ability to better capture complex, non-linear interactions between characteristics, offering a more nuanced approach compared to

traditional, linear models. However, these models still fundamentally rely on marginal distributions of single characteristics.^{224–227}

As such, despite these recent advancements in lead exposure mapping, newer machine learning models share several limitations with traditional quantile classification models given their shared reliance on marginal data. First, these models both suffer from the ecological fallacy, as they make predictions based on marginal data at the area level and can mistakenly assume that the relationships observed at the area level also apply at the individual level. Second, the aggregation process used to produce marginal data smooths over critical individual and spatial variability, potentially masking small but significant variations in residential lead exposure hotspots driven by differences in population density and diversity. Prior studies evaluating marginal distribution models have found that they tend to overrepresent urban residential lead exposure hotspots while consistently underrepresenting rural residential lead exposure hotspots.²²¹ This under-detection in rural areas is often attributed to the inability of marginal distribution models to capture localized risks within sparsely-populated areas. Our analysis confirmed this observation, with significant under-detection in states with many rural areas like Nebraska and Iowa, but also revealed significant under-detection in states with many urban areas like New York, Illinois, Massachusetts, and California. This pattern suggests that marginal distribution models may also struggle to capture the compounding risks that arise from complex interactions between population characteristics and housing characteristics in diverse, densely-populated areas. In contrast, marginal distribution models performed relatively well in less diverse states like Vermont, Nevada, Maine, Alaska, and those across the South, where the underlying assumptions of marginal distribution approaches—including consistent co-variation between population characteristics and housing characteristics and limited variability within those characteristics—are more likely to hold true. Third, marginal distribution models place undue reliance on housing age as a proxy for residential lead exposure, which fails to

account for housing quality.²²⁸ Lastly, because BLL testing is often targeted to areas based on residential lead exposure maps derived from these marginal distribution models, when these BLL data are in turn used to train models to predict residential lead exposure hotspots, they may confirm and amplify these existing biases in the data. By reinforcing these biases, this feedback loop may lead to under-testing of areas that were not initially identified as high-risk and/or over-testing of areas that were initially identified as high-risk, and ultimately the misallocation of resources.

These limitations are particularly consequential for minoritized populations, who may not only be at a heightened risk of being misrepresented or overlooked by traditional marginal distribution approaches, but also often lack the resources to mitigate lead exposure and its adverse health effects. We use our findings on four groups—the uninsured, Native American, Black, and White populations—as case studies illustrating the various limitations of current marginal distribution approaches and the advantages of our novel joint distribution approach in uncovering nuanced, context-specific vulnerabilities. By leveraging our synthetic U.S. population, we present accompanying maps that visualize the precise geographic locations of these vulnerabilities, offering unique insights into where specific subpopulations facing compounding residential lead exposure risks are concentrated and therefore where policies aiming to mitigate these risks should focus.

Uninsured populations.

Uninsured children at “high risk” of lead exposure are particularly susceptible to under-detection by current residential lead exposure maps, where every 1% increase in the proportion of uninsured populations in a census tract increased the odds of that tract going undetected as “high-risk” by 73%. This effect may have been observed because uninsured children tend to be concentrated in rural areas across the U.S.—as shown in **Figure 9**—where economic constraints and geographic isolation contribute to higher rates of

uninsurance.²²⁹ Unlike urban areas, rural economies are largely driven by industries like agriculture, manufacturing, and mining, which often lack employer-sponsored health coverage.^{229–231} Additionally, self-employment and small business ownership are common, but the cost of private health insurance is often prohibitively high.^{229–231} In states that have not expanded Medicaid—programs that require BLL testing at 12 and 24 months—uninsured rural populations have even fewer pathways to health coverage.^{229–231} These economic constraints, coupled with limited access to healthcare services in areas that are geographically isolated, leave uninsured rural populations with few opportunities for timely BLL testing, diagnosis, and treatment.²²⁹

Compounding their risk of lead exposure and barriers to lead mitigation is that uninsured rural children are 62% more likely to reside in older homes, and a large body of literature suggests that these homes are also substandard.^{231,232} Driven by limited infrastructure investment, longstanding housing shortages have led to a reliance on an aging housing stock in rural areas, many of which were built before lead regulations were enacted.^{231–236} Without robust federal support for new affordable housing or comprehensive lead abatement programs, these families often remain in aging homes with lead-contaminated paint, plumbing, and soil that continually threaten the health and safety of their children.^{231,232}

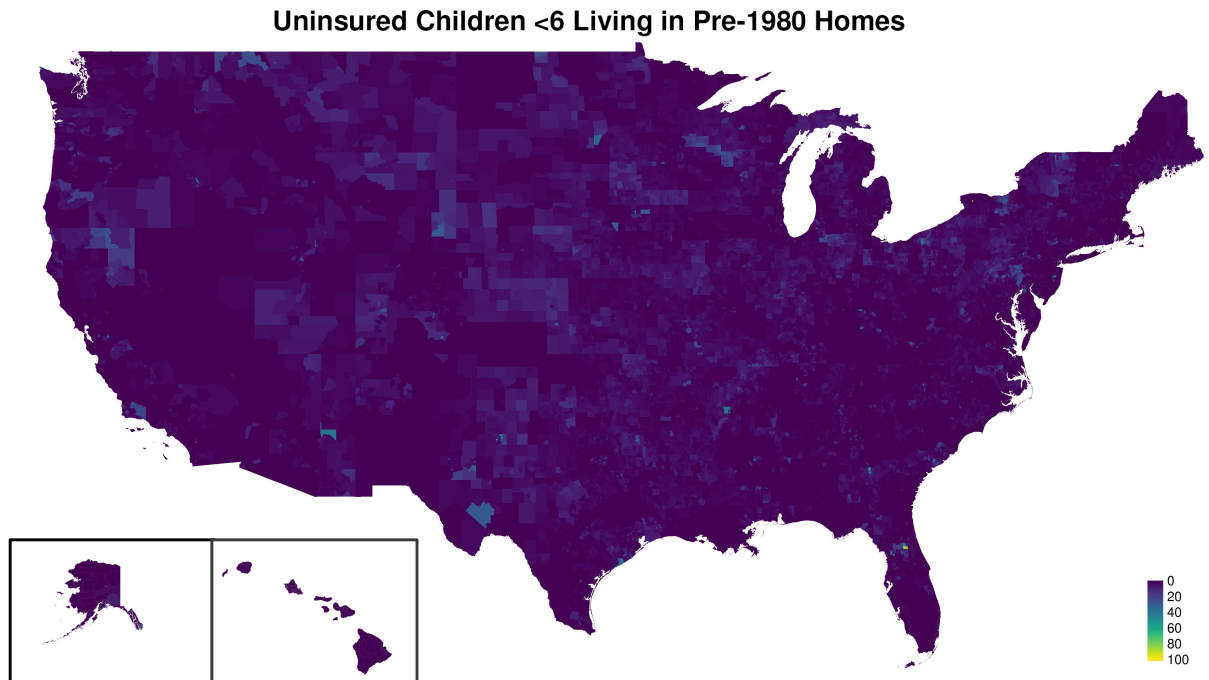


Figure 9. Concentration of uninsured children under the age of six living in homes built before 1980 at the census tract level across the U.S.

Native American populations.

Native American children at “high risk” of lead exposure face unique challenges that make them particularly susceptible to under-detection by current residential lead exposure maps, where every 1% increase in the proportion of Native American populations in a census tract increased the odds of that tract going undetected as “high-risk” by 18%. As seen in **Figure 10** clusters of Native American children living in pre-1980 homes are concentrated in regions such as the Navajo Nation in the Southwest, the Lakota reservations in the Northern Plains, and the Pueblo reservations in New Mexico.²³⁷ This distribution is shaped by the U.S. government’s trust relationship with tribal nations that has resulted in the creation of reservations²³⁸ that are often characterized by low population densities spanning large census tracts.²²⁹ While these reservations may be sparsely populated overall, Native

American communities are often clustered within reservations and Native American families are often even further clustered within multigenerational homes,²³⁹ forming pockets of risk that are easily diluted by traditional residential lead exposure maps. This geographic isolation not only may limit the likelihood of being targeted for BLL testing but also restricts access to timely diagnosis and treatment when exposure does occur.

While Native American children are 6% less likely to reside in homes built before 1980 compared to their White counterparts, housing age alone does not reflect the quality, condition, or autonomy their families have over their homes. Federal housing programs such as the Indian Housing Block Grant (IHBG), established by the Indian Housing Authority's (IHA) Native American Housing Assistance and Self-Determination Act (NAHASDA) in 1996, have focused on addressing housing insecurity for Native American tribes by constructing new homes.²⁴⁰ Despite this newer housing stock, these homes frequently suffer from an aging infrastructure, poor construction materials, and inadequate maintenance due to chronic underfunding.²⁴⁰ Furthermore, the unique land ownership structure of reservations—where land is held in trust by the federal government—limits individual autonomy over home repairs or renovations, posing significant challenges for Native American residents seeking to address lead hazards in their own homes.²⁴⁰

Also not captured by either approach is the environmental context in which many reservations are located. Historically, the federal government allocated lands to Native American tribes that were deemed undesirable for mainstream economic development, often placing reservations near or on mining operations, industrial sites, and waste zones that emit toxic pollutants.^{241–244} Proximity to these sites has been associated with higher risk of lead-contaminated air, water, soil, and dust, which tend to accumulate in homes, posing additional risks to Native American residents.^{243–246}

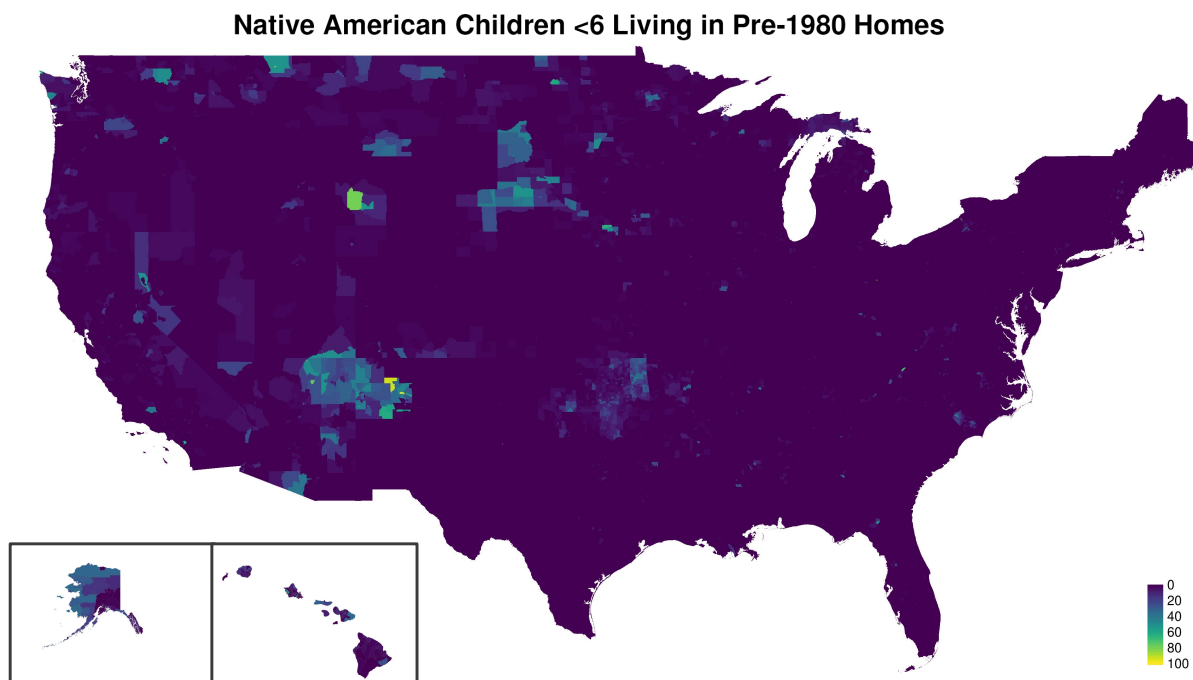


Figure 10. Concentration of Native American children under the age of six living in homes built before 1980 at the census tract level across the U.S.

Black populations.

Interestingly, although our findings reveal Black children are less likely than their White counterparts to live in older homes, an extensive body of literature has established that they are significantly more likely to have elevated BLLs^{198,199,247–249}—a paradox underscoring the importance of considering broader social, economic, infrastructural, and environmental contexts when using proxies to quantify lead exposure. While federal urban renewal programs and the 1968 Fair Housing Act facilitated the construction of newer apartment complexes and affordable housing developments in urban areas across the South and Northeast,^{250,251} Black families have largely remained segregated into urban neighborhoods facing greater environmental risks due to structural racism, as shown in **Figure 11.**²⁵² Long-standing practices like redlining, discriminatory lending, and urban

renewal policies funneled Black communities into urban areas deemed undesirable—often situated near highways, industrial sites, or waste zones—which have been significant sources of both legacy and ongoing lead contamination.^{33,36,45} As a result, Black families in these newer dwellings may still face risks of lead exposure not only from aging infrastructure, poor construction, and inadequate maintenance, but also from their surrounding environmental context,^{253–255} similar to the risks faced by Native American families.²⁴⁴ Compounding these risks is that Black families often encounter barriers to lead abatement, lacking the necessary resources to effectively mitigate exposure.²⁵⁶

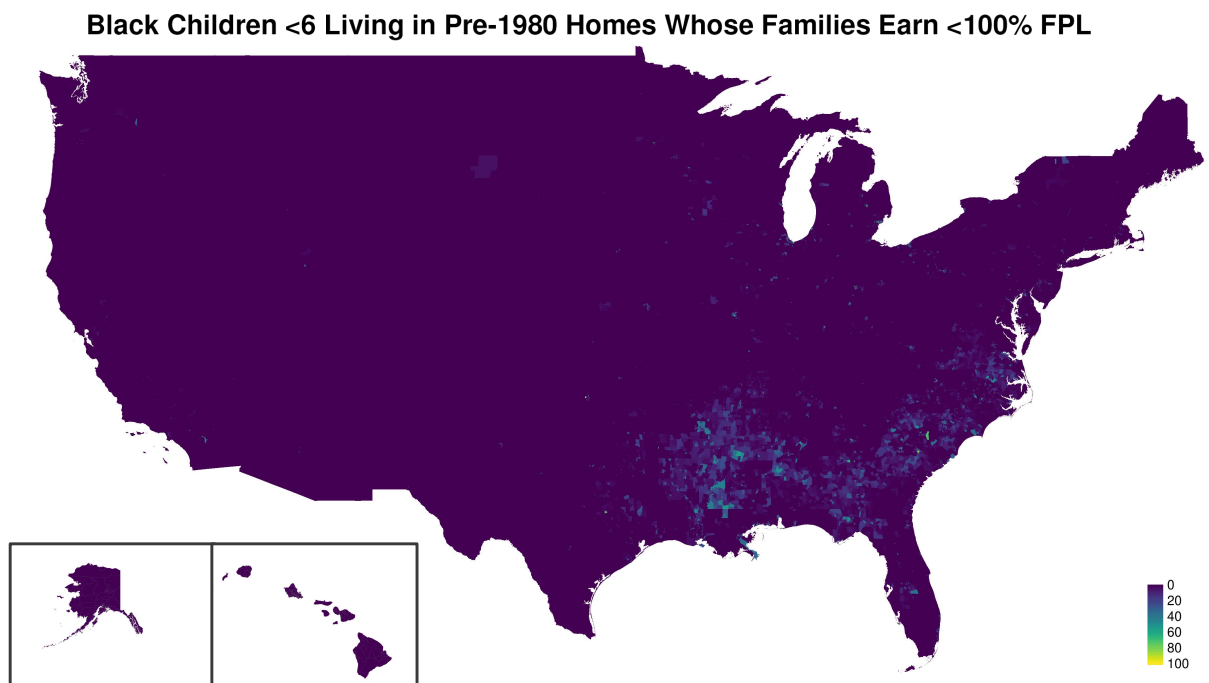


Figure 11. Concentration of impoverished Black children under the age of six living in homes built before 1980 at the census tract level across the U.S.

White populations.

Conversely, White children were more likely to live in older homes than any other racialized group, yet they are significantly less likely to have elevated BLLs. This discrepancy

may be attributed to generational homeownership, with White families often passing homes down through generations, along with greater access to financial resources for lead abatement. These factors may allow White families to reside in older homes while effectively mitigating the risks of lead-based paint exposure. However, it remains critical to identify and allocate resources to White children living in older homes who have gone undetected or lack the means to mitigate these risks, as shown in **Figure 12**.

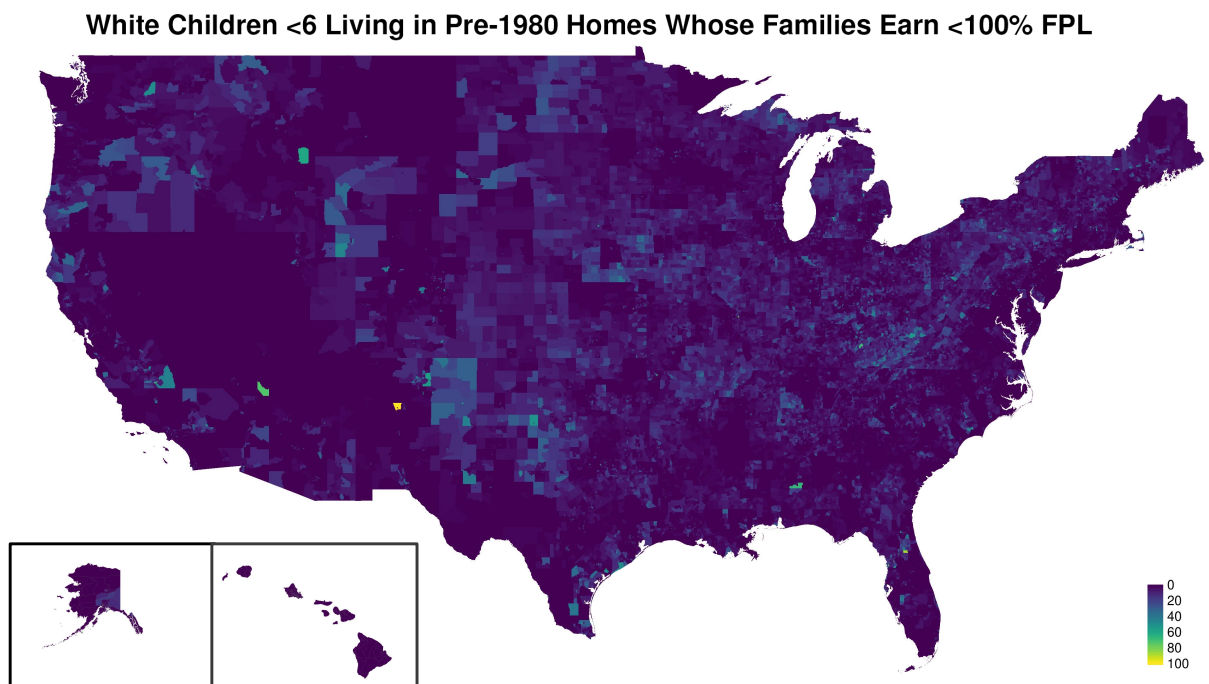


Figure 12. Concentration of impoverished White children under the age of six living in homes built before 1980 at the census tract level across the U.S.

5.1. Strengths.

Current approaches to identifying residential lead exposure hotspots often hide “high-risk” areas because they rely on marginal distributions of single characteristics. This represents a significant limitation, as these methods not only underestimate the full extent of risk, but also fail to account for the full nature of risk arising from compounded social, economic, infrastructural, and

environmental burdens faced by those living in hidden “high-risk” areas. In contrast, our novel approach uncovers those high-risk areas by capturing the joint distributions of multiple characteristics. This enables us to pinpoint populations living in systematically under-resourced areas—such as uninsured children living in rural areas, Native American children living in reservations, and low-income Black children living in urban areas across the South and Northeast—who require targeted lead abatement efforts. Furthermore, by preserving the full variability in the data, our novel model provides a more nuanced understanding of the gradient of risk, rather than relying on the broad quantile-based categorizations of traditional models. Relatedly, our novel model provides flexibility in variable selection and manipulation. For instance, age is treated as a continuous variable in the synthetic U.S. population, offering the ability to define any age range or even isolate specific ages rather than being confined to the ACS’s rigid bands of 0-4 years, 5-9 years, 10-14 years, and so forth. This allows us to tailor analyses to the most at-risk subpopulations—such as children under the age of six (an age group identified by the CDC as especially vulnerable to lead exposure)—but also extends to exploring any other ages or age groups for a more nuanced understanding of other at-risk subpopulations. Additionally, our synthetic U.S. population grants access to a more comprehensive array of target variables that are not otherwise available in traditional marginal datasets, enabling the incorporation of nuanced characteristics like socioeconomic indicators, insurance status, housing conditions, work conditions, and time activity patterns, facilitating a comprehensive and context-sensitive evaluation of lead exposure risks.

5.2. Limitations.

Despite these strengths, our novel joint distribution approach is not without limitations. First, our model is inherently subject to the same limitations in data quality, modeling assumptions,

and underlying structure as the synthetic U.S. population itself, which are outlined in detail in Parikh et al. (2024). However, extensive validation of the synthetic U.S. population has demonstrated that it performs well in reflecting the real US population at the census tract level across several key social, economic, infrastructural, and environmental characteristics. Second, the case study presented here is based on a simplified scenario that focuses on the joint effects of age and housing age as a proof-of-concept for how our synthetic U.S. population can improve the precision of current residential lead exposure approaches. While this provides a useful demonstration, it is not a direct one-to-one comparison with existing residential lead exposure maps that often layer in additional marginal distributions such as poverty, race, or other characteristics. A future study incorporating a wider range of variables—validated against actual BLL data from states like Massachusetts that mandate universal BLL testing—could further illustrate the benefits and applications of our novel joint distribution approach. Additionally, while we used housing age as a residential risk proxy to align our model with current mapping methods, we acknowledge that this metric alone is insufficient in capturing residential lead exposures. More comprehensive indicators of residential lead exposure would also account for factors like housing quality, infrastructure condition, and neighborhood environmental risks including the presence of lead-based pipes and lead-emitting sites. Relatedly, this study only considers young children given their heightened susceptibility to lead exposure. However, emerging research suggests that other groups such as young women and pregnant individuals may also face significant risks from lead exposure, and broadening the scope of the analysis to consider these additional vulnerable groups would enhance the model’s utility and relevance for more inclusive public health interventions.

6. CONCLUSIONS.

In response to the federal government’s call for improved methods to map residential lead exposure hotspots, we leveraged joint distributions from our novel synthetic U.S. population to identify high-risk areas and compared these results to traditional methods relying on marginal distributions from the ACS. Specifically, our census-tract-level analysis revealed that the majority of at-risk areas identified through our novel joint distribution approach were not captured by traditional marginal distribution approach, representing millions of at-risk young children who may have gone undetected in traditional risk assessments and mitigation efforts. Furthermore, areas with high concentrations of White, Native American, and uninsured individuals were systematically more likely to go undetected by traditional approaches. Additional individual-level analyses showed that uninsured and impoverished children were more likely to reside in older homes, further underscoring the limitations of existing methods and the need for more comprehensive, context-sensitive approaches.

Our study demonstrated that our synthetic U.S. population can provide a powerful tool for identifying both who is at risk and where those risks are concentrated, and underscores the critical need for more comprehensive and context-sensitive approaches to identifying residential lead exposure hotspots to better target at-risk communities for intervention. By integrating joint social, economic, infrastructural, and environmental characteristics through our synthetic U.S. population, we have taken an important step toward more accurately capturing the multifaceted, multilevel vulnerabilities faced by marginalized communities—vulnerabilities that are often overlooked by current residential lead exposure maps. This approach not only enhances the precision of identifying pockets of risk, but also provides a pathway for developing interventions tailored to the specific needs and contexts of these communities, all of which are barriers associated with other solutions like mandated universal BLL testing.

The implications of our findings extend beyond the technical advancements in risk mapping; they point to a broader call for policy shifts, resource allocation, and community engagement to break the cycle of vulnerability to lead exposures. Specifically, there is an urgent need to address gaps in current methods for identifying residential lead exposure hotspots for the vulnerable subpopulations that have thus far gone undetected, unprotected, and underserved. Future research should build on our work by refining risk models to include more comprehensive indicators of housing, neighborhood, and infrastructural quality that are relevant to residential lead exposure risks, such as HUD's Deteriorating Paint Index,²⁵⁷ the Natural Resources Defense Council's (NRDC) map of aging lead-based water pipes,²⁵⁸ contaminated outdoor soil, and proximity to highways, industrial sites, mining operations, and waste zones. Further, modeling lead exposures, BLLs, or lead-attributed health outcomes as a function of multiple characteristics could improve our understanding of individual risk profiles with geographic specificity. More advanced methods, like agent-based simulations, could also be employed to simulate the movements of individuals throughout their environments each day, offering more precise estimates of true lead exposures. Lastly, future scenario simulations could be used to evaluate the potential impact of policy changes, lead abatement interventions, and resource reallocation on reducing lead exposures in vulnerable communities. These research directions will not only enhance our capacity to identify and mitigate lead hazards, but also support the development of equitable and effective public health strategies for all communities.

CHAPTER 4: LEVERAGING A SYNTHETIC POPULATION TO IDENTIFY POPULATIONS UNPROTECTED AGAINST HEAT STRESS ACROSS THE UNITED STATES

Shivani Parikh^{a,b}, Lauren Ferguson^a, Brent A. Coull^a, John D. Spengler^a, Joseph G. Allen^a

^a Harvard T.H. Chan School of Public Health, 401 Park Drive, 4th Floor West, Boston, MA,
02215, USA

^b Harvard Graduate School of Arts and Sciences, 1350 Massachusetts Avenue, Cambridge, MA,
02138, USA

1. ABSTRACT.

As climate change intensifies, heat stress has become a major public health concern across the United States (U.S.). Residential air conditioning (AC) is critical for mitigating heat-related health and economic burdens, yet disparities in residential AC access continue to place vulnerable populations at risk. Traditional approaches to assess these disparities often rely on residential AC ownership data, reported either as individual-level microdata at coarse spatial resolutions or as geographically-aggregated data at finer spatial resolutions, which fails to capture the financial and logistical barriers to residential AC use, as well as important spatial and individual variability in residential AC access.

To overcome these limitations, we trained a gradient boosting model on 2020 microdata from the Residential Energy Consumption Survey (RECS) to predict residential AC access—defined here as the ability to own, install, maintain, and operate AC systems—as a function of age, sex, race, ethnicity, household income, housing tenure, occupancy, building type, building age, cooling degree days (CDD), and state. We then applied this model to 2018-2022 microdata from a synthetic U.S. population comprising 330 million agents designed to reflect the real U.S. population at the census tract level, to assign each agent an individual probability of residential AC access based on these same predictive characteristics. This approach allowed us to downscale individual residential AC access data that traditionally only provides spatial information at the state level to the census tract level, enabling a fine-grained analysis of how compounding social, economic, infrastructural, and environmental characteristics shape residential AC access across diverse U.S. contexts.

We found stark contrasts between our findings and those of existing research: while studies report a median residential AC ownership probability of 97% with values ranging from 15-100%, our model estimated a median residential AC access probability of only 80% with values ranging from 3-97%, discrepancies which we attribute to our unique inclusion of both AC ownership and

AC use in our measurement of AC access. Notably, state, household income, and building age emerged as among the strongest predictors of AC access in our model. Limited residential AC access was most pronounced in the West, Northwest, and Northeast states, but significant intrastate variability persisted across states within all U.S. regions, underscoring the need for assessments that account for spatial variations in state policies and programs, energy infrastructures and costs, building practices and standards, environmental conditions, population density, and population composition. Household income showed a pronounced gradient, where lower-income households were up to 24% more likely to experience limited residential AC access than higher-income households. A similar pattern emerged for building age, with those residing in the oldest homes being up to 13% more likely to experience limited residential AC access than those in newer homes. Additional disparities were found across race, ethnicity, housing tenure, occupancy, building type, and CDDs.

We envision that federal, state, and local agencies could use this data-driven tool to identify, prioritize, and allocate cooling resources to populations and places most at risk, advancing climate resilience and promoting climate equity.

2. INTRODUCTION.

With the global climate crisis rapidly intensifying over the past century, protecting the population against heat stress has emerged as one of today's most pressing public health challenges. Exposure to heat stress is directly linked to a range of acute and chronic conditions—from headaches, dizziness, nausea, impaired focus, heat exhaustion, and heat stroke to cardiovascular, respiratory, renal, and neurological diseases—all of which contribute to higher risk of morbidity and mortality compared to that of any other weather-related hazard.^{259–265} These health impacts also carry

significant economic consequences, resulting in over \$1 billion in healthcare costs and an additional \$100 billion in productivity losses each year in the U.S. alone.^{266–268}

Given that people spend the majority of their time indoors—particularly within their homes—the conditions of these environments profoundly shape their exposures to heat stress.²⁶⁹ Among available options, air conditioning (AC) stands out as the most widespread and effective cooling solution for maintaining safe indoor temperatures in these environments, and has been shown to reduce the risk of heat-related illnesses.^{270–273} However, while approximately 90% of U.S. households own an AC unit,²⁷⁴ ownership alone does not guarantee its use. True access to AC as a cooling solution is often constrained not only by the cost of the equipment but also by the expenses associated with installation, maintenance, and the ongoing energy required for operation.^{275–278} This challenge reflects a broader energy burden affecting a significant portion of the U.S. population, defined as spending more than 6% of household income on energy costs.²⁷⁹ Households facing this burden are often forced to make difficult trade-offs between maintaining safe indoor temperatures and meeting other essential needs.²⁷⁹ Currently, more than 30% of U.S. households are classified as energy-burdened, and this challenge is only expected to grow as climate extremes intensify, the housing stock ages, and energy costs rise.^{280–284}

These burdens are not distributed equally across the population. A long history of entrenched structural discrimination in the U.S. has segregated economically disadvantaged and racially- and ethnically- minoritized communities into neighborhoods with disproportionately high levels of heat stress.^{285–288} Many homes in these neighborhoods are aging, often lacking proper insulation and energy-efficient cooling systems, which makes maintaining safe indoor temperatures both challenging and costly.^{278,279,289–294} Moreover, these marginalized communities often face significant barriers to managing heat-related illness due to unaffordable healthcare coverage, fewer healthcare facilities, and substandard healthcare options.^{295–297} This cycle of vulnerability—where

those most exposed to heat stress are least able to protect against it—is particularly severe for children and the elderly belonging to these communities, who also face increased physiological susceptibility to extreme temperatures due to reduced thermoregulatory capacity.^{297–299}

Therefore, understanding which populations face limited access to cooling and pinpointing where they live represent essential steps in designing and implementing accessible cooling interventions that reduce the compounding burdens associated with heat stress. However, research in this area is limited due to two key methodological limitations. First, few datasets capture residential AC access, forcing most researchers to rely on residential AC ownership as a sufficient proxy for residential AC access.^{291,300–308} This reliance overlooks details that are essential to determining one’s ability to maintain safe indoor temperatures during peak heat season including whether households can afford to own, install, maintain, and operate AC systems as described above. Second, data on residential AC ownership is typically available in one of two forms to protect respondent privacy, each with distinct limitations: individual-level microdata at coarse spatial resolutions (state or national level), or as geographically-aggregated data at finer spatial resolutions (census block, census tract, or county level).^{309,310} Individual-level microdata at coarse spatial resolutions obscures important spatial variability, where different neighborhoods even within the same county can differ substantially in demographic makeup, housing quality, resource availability, energy costs, and climate conditions—characteristics influencing residents’ ability and need to access AC. Geographically-aggregated data at fine spatial resolutions, on the other hand, masks important individual variability on how these key social, economic, infrastructural, and environmental characteristics interact within specific subpopulations to create varying degrees of vulnerability in limited AC access.

The U.S. Energy Information Administration (EIA) made a significant contribution to energy justice research with its release of the 2020 Residential Energy Consumption Survey (RECS),

which, for the first time, provided national household-level microdata on the ability to own, install, maintain, and operate AC systems.³⁰⁹ However, since this dataset only provides spatial information on the state where each surveyed household is located, its utility for planning cooling interventions to mitigate heat stress at the local level is limited. To bridge this gap, we harness the power of a novel synthetic U.S. population we previously generated, which provides realistic individual-level microdata on key social, economic, infrastructural, and environmental characteristics along with spatial information on the census tract where each synthetic agent is located. Specifically, we used machine learning techniques to model residential AC access as a joint function of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, Cooling Degree Days (CDD), and state using RECS data. We then applied this model to predict residential AC access for each of the 330 million agents within the synthetic U.S. population, assessing the state-, tract-, and individual-level characteristics that produce differential vulnerabilities to residential AC access. By integrating microdata from RECS and our synthetic U.S. population, this study represents the first to map individual residential AC access probabilities across all U.S. census tracts, providing actionable insights into which populations face limited access to cooling and where they live.

3. METHODS.

We developed a novel approach that combines microdata from the RECS and synthetic U.S. population datasets to model individual probabilities of residential air conditioning (AC) access at a fine spatial resolution across the U.S. RECS provides household-level microdata on AC access—defined here as the ability to own, install, maintain, and operate AC systems—along with several characteristics known to influence AC access including age, sex, race, ethnicity, household income,

housing tenure, housing occupancy, building type, building age, CDD, and state with spatial information at the state level. While the synthetic U.S. population lacks direct data on AC access, it contains individual-level microdata on these same predictive characteristics with spatial information at the census tract level, making it an ideal candidate for downscaling AC access estimates to a fine spatial resolution across the U.S. using predictive modeling. To implement this approach, we followed six key steps (**Figure 13**):

- (1) Selecting the target variable for residential AC access from the RECS dataset;
- (2) Standardizing the predictive variables across the RECS and synthetic U.S. population datasets;
- (3) Harmonizing the underlying structures of the RECS and synthetic U.S. population datasets;
- (4) Balancing the two residential AC access classes;
- (5) Training and validating a gradient boosting model on RECS data to assign individual probabilities of residential AC access with spatial information provided at the state level;
- (6) Applying this gradient boosting model to the synthetic U.S. population dataset to assign individual probabilities of residential AC access to each of the 330 million synthetic agents with spatial information provided at the census tract level.

This novel approach enabled us to analyze how residential AC access is shaped by the interplay of social, economic, infrastructural, and environmental characteristics at a fine spatial resolution across diverse U.S. contexts (described in subsequent sections). All analyses were conducted using R version 4.3.3.

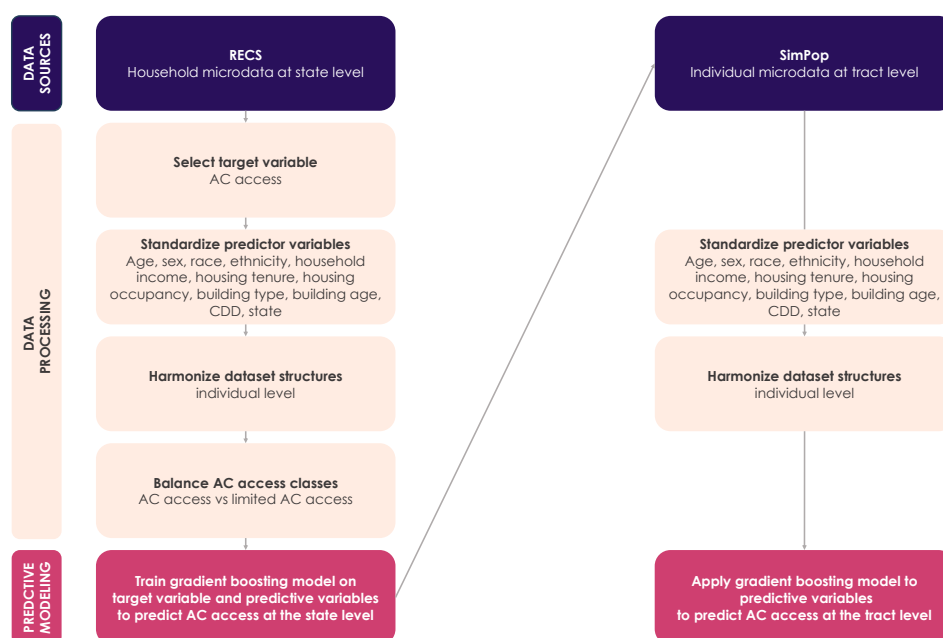


Figure 13. Methodological framework.

3.1. Data sources.

3.1.1. RECS dataset.

The Residential Energy Consumption Survey (RECS)—administered by the U.S. Energy Information Administration (EIA)—is designed to estimate energy consumption, expenditures, and characteristics of occupied housing units across the U.S.³¹¹ In 2020, RECS introduced household-level microdata collected from 18,496 nationally representative households via self-administered web and paper questionnaires across all 50 states and the District of Columbia with spatial information specifying the state where each household is located.³¹¹ The RECS includes a voluntary Household Survey, which collects data from households on energy consumption, and the mandatory Energy Supplier Survey (ESS), which gathers data from energy suppliers on actual consumption and

expenditures for electricity, natural gas, and other fuels.³¹¹ A comprehensive nonresponse bias analysis was conducted to ensure data quality, and no significant concerns were found, indicating that the final weighted estimates closely represent the target U.S. population.³¹¹

3.1.2. Synthetic U.S. population dataset.

In a previous study, we generated a novel synthetic U.S. population consisting of 330,587,560 synthetic agents, the methods for which are described in detail in Parikh et al. (2024). This dataset provides realistic individual-level microdata on key social, economic, infrastructural, and environmental characteristics along with spatial information specifying the census tract where each synthetic agent is located. The synthetic U.S. population was designed to mirror real-world joint distributions at a fine spatial resolution across the entire U.S.

Briefly, our analysis integrated two primary datasets from the U.S. Census Bureau’s American Community Survey: the 2022 5-Year ACS, which provides comprehensive geographically-aggregated data on the U.S. population at the census tract level, and the 2022 1-Year ACS Public Use Microdata Sample (PUMS), a representative subset offering detailed individual- and household-level microdata on approximately 1% of the U.S. population at the Public Use Microdata Area (PUMA) level. Together, these datasets allowed us to uniquely model joint distributions of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel at the census tract level using the best available data.

We employed a static spatial microsimulation approach using Iterative Proportional Fitting (IPF), a deterministic reweighting method known for its ability to efficiently generate consistent estimates with each run.^{54,84–87,94} We first categorized the individual-level microdata by converting categorical variables into binary dummy variables (0/1)—an alignment that was crucial for ensuring

that the individual-level microdata could be constrained to the geographically-aggregated data in subsequent steps. Initial weights were set to 1 so that each individual started with equal influence, and during the IPF process, these weights were iteratively adjusted within each PUMA using the `ipfp` package in R.⁵⁴ The process ran for up to 50 iterations or until convergence was reached at an acceptable tolerance threshold of $1.5\text{e-}8$ between the adjusted weights and known constraints. We converted the resulting fractional weights from the IPF process into integers through the pseudorandom Truncate, Replicate, Sample (TRS) method.⁹¹ Finally, we expanded the integerized data by replicating individuals according to their integer weights.

The resulting synthetic U.S. population underwent extensive internal and external validation to ensure its accuracy and applicability for further public health analyses. For internal validation, we achieved a Pearson's R of 0.99 across census tracts and a standard absolute error (SAE) below 0.1 for 97% of tracts. We externally validated each demographic variable against the 2020 U.S. Decennial Census and each housing variable against the 2020 RECS dataset using the Kullback-Leibler divergence statistic, achieving means of <0.08 across all variables, indicating very little information loss.

3.2. Data processing.

3.2.1. Selecting the target variable for AC access from the RECS dataset.

Our analysis aimed to assess AC access—not only whether households owned an AC system but also whether they could afford to use it consistently—as our target variable.

To capture this, we selected “no AC days” as our target variable from the RECS dataset, which records the number of days households went without AC over the year. This variable, originally continuous (0–365), displayed a bimodal distribution with most households clustered

around either 0 or 365 days and few intermediate values between 0 and 365 days. This extreme clustering can pose a significant challenge for predictive models, particularly for gradient boosting, which generally perform best with either binary data that has balanced classes or continuous data that has a broad distribution across its range. Gradient boosting models build successive decision trees where each new tree attempts to correct errors from previous iteration to improve its predictive accuracy. However, when data points are concentrated at the extremes and limited across its full range of values, the model is forced to primarily correct errors within these dominant clusters and can end up “memorizing” the patterns at 0 or 365 days rather than learning generalizable relationships between 0 and 365 days. Overfitting on the dominant clusters in this way ultimately reduces the model’s sensitivity to subtler variations that might be associated with intermediate levels of AC access.

To address this challenge, we transformed “no AC days” into a binary target variable: 0 for households with no days without AC and 1 for households with any days without AC over the year. This transformation allowed us to accomplish our goal of distinguishing between households that could afford consistent AC access and those who could not for any reason.

We then cross-referenced this binary target variable against other indicators available within the RECS dataset that signal different dimensions of limited AC access, including whether households reported having an AC system, whether households were unable to use their AC systems due to equipment failure they could not afford to repair or replace, and whether households were disconnected from electricity necessary to operate their AC systems due to energy bills they could not afford to pay. We found that all households flagged by the binary target variable as having limited AC access were also identified by at least one of these other indicators of limited AC access. Furthermore, no households flagged by the other indicators were missed by the binary target variable, indicating that it captured all measured dimensions of limited AC access within the RECS

dataset. To quantify the level of agreement between the binary target variable and these other indicators of AC access, we calculated Cohen's Kappa, a statistical measure of inter-rater reliability that accounts for agreement occurring by chance. The results revealed a strong positive association between the binary target variable and AC ownership status ($k = 0.84$) and modest to low positive associations with AC repair affordability and AC electricity affordability ($k = 0.35$ and $k = 0.06$, respectively).

3.2.2. Standardizing the predictive variables across the RECS and synthetic U.S. population datasets.

To downscale this binary target variable, “no AC days,” to a fine spatial resolution, we implemented a gradient boosting model to identify patterns of AC access based on a set of predictive individual-level characteristics in the RECS dataset (training data) and applied these learned patterns based on this same set of predictive individual-level characteristics in the synthetic U.S. population dataset (application data).

A critical step in implementing this approach was selecting and standardizing the predictive variables common to both the RECS and synthetic U.S. population datasets. This is because gradient boosting models are highly sensitive to the consistency of input data; if variable definitions, coding, or formats differ between the training and application phases, the model may misinterpret the relationships it has learned when transferring them from the training data to the application data, leading to biased or unreliable predictions. We therefore selected age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDD, and state as our predictive variables and standardized them across datasets to enable our model to consistently interpret relationships across the training and application phases. This selection was based on

research highlighting how these social, economic, infrastructural, and environmental characteristics influence AC access across diverse U.S. contexts.

3.2.3. Harmonizing the underlying structures of the RECS and synthetic U.S. population datasets.

An equally critical step in implementing this approach was harmonizing the underlying structures of the RECS dataset (which reports data on households) with the synthetic U.S. population dataset (which contains data on individuals), given that gradient boosting models require consistency in input data to make accurate predictions, as detailed in the previous section.

To achieve this consistency, we developed a custom function to extrapolate householder characteristics to all household members in the RECS dataset, treating them as individual data points. We initially replicated the characteristics of the original householder across all household members to reflect the reported number of people in each household, however, we then tailored these characteristics based on different criteria for different variables.

Household-level variables—including household income, housing tenure, housing occupancy, building type, building age, CDD, state, and AC access—were uniformly applied to each household member based on the householder’s data since these characteristics pertain to the household as a whole. Race and ethnicity variables were also uniformly assigned to each household member based on the householder’s data, aligning with national trends indicating that over 90% of U.S. households consist of individuals of the same race and ethnicity, which minimizes assumptions about intra-household diversity.

For the age variable, the RECS dataset provided the number of occupants in three broad age groups: children (0–17), adults (18–64), and elderly individuals (65+). We assigned each household member to an age group based on this information.

For the sex variable we leveraged the 2022 American Community Survey (ACS), the largest annual survey of the U.S. population conducted by the U.S. Census Bureau that provides state-specific estimates of marginal sex distributions. The ACS’s large sample size and state-level granularity allowed us to probabilistically assign these characteristics to household members in a way that preserved realistic demographic patterns across states. We assigned each household member a sex based on state-specific sex distributions to force RECS to reflect real-world variations in sex ratios at the state level. **Table 11** details the final dataset structures.

Table 11. Aligning the structures of the RECS and synthetic U.S. population datasets to facilitate predictive modeling.

Variable	Original RECS coding	Original synthetic U.S. population coding	Standardized coding
Level	Household	Individual	Individual
Age	0-17 18-64 65+	Years	0-17 18-64 65+
Sex	Male Female	Male Female	Male Female
Race	White Black Asian Native Other	White Black Asian Native Other	White Black Asian Native Other
Ethnicity	Hispanic Non-Hispanic	Hispanic Non-Hispanic	Hispanic Non-Hispanic
Household income	> \$5000 \$7500-\$9999 \$10000-\$12499 \$12500-\$19999 \$20000-\$24999 \$25000-\$29999	Dollars	> \$5000 \$7500-\$9999 \$10000-\$12499 \$12500-\$19999 \$20000-\$24999 \$25000-\$29999

	\$30000-\$34999 \$35000-\$39999 \$40000-\$49999 \$50000-\$59999 \$60000-\$74999 \$75000-\$99999 \$100000-\$149999 > \$150000		\$30000-\$34999 \$35000-\$39999 \$40000-\$49999 \$50000-\$59999 \$60000-\$74999 \$75000-\$99999 \$100000-\$149999 > \$150000
Housing tenure	Rent Own Occupy without payment of rent	Rent Own Occupy without payment of rent	Rent Own (including occupy without payment of rent)
Housing occupancy	Number of people	Number of people	Number of people
Building type	Mobile home Apartment (2-4) Apartment (>5) Single family home (attached) Single family home (detached)	Mobile home Boat, RV, van Apartment (2) Apartment (3-4) Apartment (5-9) Apartment (10-19) Apartment (20-49) Apartment (>50) Single family home (attached) Single family home (detached)	Mobile home (including boat, RV, van) Apartment (including all apartment types) Single family home (including all single family home types)
Building age	Before 1950 1950-1959 1960-1969 1970-1979 1980-1989 1990-1999 2000-2009 2010-2015 After 2015	Before 1940 1940-1949 1950-1959 1960-1969 1970-1979 1980-1989 1990-1999 2000-2009 2010-2019 2020 2021 2022	Before 1950 1950-1959 1960-1969 1970-1979 1980-1989 1990-1999 2000-2009 After 2010
CDD	CDDs (defined as the number of degrees that a day's average temperature is above a base temperature of 65°F, where the average temperature for each household was assigned from their nearest	CDDs (defined as the number of degrees that a day's average temperature is above a base temperature of 65°F, where the CDDs for each individual was assigned based on their census tract designation	CDDs

	weather stations extracted from the National Centers for Environmental Information's (NCEI) Climate Data Online (CDO) database for the year 2020 ³¹¹)	extracted from the Office of Energy Efficiency and Renewable Energy National Renewable Energy Laboratory's (NREL) State and Local Planning for Energy (SLOPE) tool under the Representative Concentration Pathway (RCP4.5) scenario for the year 202 ³¹²)	
State	State (serving as a proxy to capture state-level differences in costs, policies, programs, practices, building standards, and energy infrastructures related to cooling needs)	State (serving as a proxy to capture state-level differences in costs, policies, programs, practices, building standards, and energy infrastructures related to cooling needs)	State

3.3.3. Balancing the two AC access classes.

Our final data processing step before training the gradient boosting model involved balancing the two AC access classes in the RECS dataset. The initial distribution of classes was highly imbalanced, with approximately 82.8% of households with consistent AC access and only 17.2% with limited AC access. Such class imbalances can negatively impact the performance of predictive models such as gradient boosting which may overfit to the majority class while under-representing the minority class.

To mitigate this imbalance, we applied the Synthetic Minority Oversampling Technique (SMOTE), a widely-favored method proposed by Chawla et al. (2002) to balance classes.³¹³ SMOTE works by identifying each instance in the minority class and locating its nearest neighbors within that class.³¹³ It then creates synthetic examples by selecting random points along the line segments between each instance and its nearest neighbors.³¹³ This interpolation-based approach allows the

model to generate diverse synthetic examples that lie between existing instances.³¹³ Unlike methods that rely on simple resampling, SMOTE avoids overfitting on repeated observations by introducing variability among the synthetic examples, making it particularly effective for maintaining diversity within the minority class.³¹³ Research has shown that balancing classes in this way can allow the model to learn meaningful patterns for both classes, thereby improving its predictive accuracy and generalizability.^{314–316}

In our analysis, we applied SMOTE with $k = 5$ nearest neighbors, following the original approach used by Chawla et al. (2002).³¹³ This setting means that each synthetic example was generated by interpolating between an instance in the minority class and one of its five closest neighbors, thereby spreading synthetic examples across the feature space of the minority class to enhance their diversity. To maintain the population’s representativeness while avoiding undue influence from synthetic instances, we preserved the original RECS dataset weights for real observations and assigned a constant weight of 1 to the synthetic examples, following best practices outlined in He and Garcia, 2009,³¹⁴ Kotsiantis et al., 2006,³¹⁵ and Fernandez et al., 2018.³¹⁶ This approach resulted in a balanced dataset, with approximately 45.4% of households classified as having limited AC access and 54.6% with consistent AC access.

3.3. Modeling.

Following data processing, we trained and validated a probabilistic classification model on the data in RECS, which enabled us to assign individual AC access probabilities to each of the 330 million synthetic agents within the synthetic U.S. population as a function of their age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDD, and state. To accomplish this, we selected gradient boosting—a robust machine learning technique

suitable for both regression and classification tasks.^{317–319} This technique works by sequentially combining multiple weak learners called decision trees to build a strong predictive model.^{317–319} A decision tree splits the data into subsets based on feature values, creating a branching structure where each internal node represents a decision and each leaf represents an outcome.^{317–319} Critically, each decision tree in gradient boosting is added to correct the errors of the previous one, with the model optimizing its predictions by minimizing a loss function through an iterative process.^{317–319} Our choice of gradient boosting for this analysis was guided by four key advantages. First, gradient boosting supports probabilistic classification, providing continuous probability estimates between 0 and 1 for each class rather than strict binary classifications of AC access or limited AC access.^{317,320} Second, gradient boosting is adept at handling various variable structures—both categorical and continuous—spanning diverse ranges.^{317,320} This adaptability is particularly beneficial in our case given the mix of social, economic, infrastructural, and environmental predictor variables in our model. Third, gradient boosting effectively captures nonlinear relationships and interactions between variables, allowing the model to identify complex patterns and dependencies that might not be apparent through simpler techniques.^{317,320} Finally, because gradient boosting allows us to control model complexity through hyperparameters like learning rate, tree depth, and the number of boosting rounds, it enables effective management of overfitting—a common challenge in other machine learning models.^{317,320}

3.3.1. Model training and validation.

We trained and validated our model on the RECS dataset, allowing us to assess the model's performance before applying it to the synthetic U.S. population dataset. To do this, we split RECS into 80% training and 20% testing sets, keeping households intact to avoid treating household

members as independent observations—which could otherwise lead to overfitting. We then applied stratified sampling to ensure that the distribution of households with and without AC access was maintained across both sets.

To optimize the model, we conducted a grid search aimed at maximizing the area under the receiver operating characteristic curve (ROC-AUC), adjusting key hyperparameters that control model complexity, learning rate, and regularization to balance predictive accuracy and generalizability.^{317,320} These parameters included the number (100, 250, 500, 750, 1000) and depth (6, 8, 10) of decision trees to regulate model complexity; the learning rate (0.01, 0.05, 0.1) to refine prediction adjustments; and regularization factors like minimum child weight (1, 5, 10), gamma (0, 1, 5), subsampling rate (0.7, 0.8, 0.9), and column sampling by tree (0.7, 0.8, 0.9), which introduce randomness and limit overfitting by shaping how each tree is built.^{317,320} Additionally, we used a 10-fold cross-validation process repeated three times to identify the optimal hyperparameter combination for maximizing ROC-AUC across various random splits. The final gradient boosting model achieved a ROC-AUC of 0.95 on the training set and 0.80 on the testing set with 500 boosting rounds, a maximum tree depth of 6, a learning rate of 0.01, a minimum child weight of 1, a gamma of 1, a subsample rate of 0.8, and a column sampling by tree of 0.8.

We then trained this validated model on the full RECS dataset to maximize data utilization before applying it to the synthetic U.S. population dataset.

3.3.2. Model application.

Finally, we applied the final gradient boosting model trained on the full RECS dataset to the synthetic U.S. population dataset, enabling us to assign individual probabilities of AC access to each of the 330 million synthetic agents within the synthetic U.S. population.

3.4. Analysis.

To understand how modeled probabilities of AC access are influenced by various characteristics across the U.S. population, we conducted a multi-level analysis. First, we identified the most influential predictors of AC access by generating a feature importance matrix using the XGBoost package in R. This matrix quantified the relative influence of each predictor based on three metrics: the frequency with which it was used to split the data (frequency), the improvement in model accuracy attributed to its use in splitting (gain), and the proportion of samples it affected across all splits (cover). Second, we compared interstate and intrastate disparities in AC access by aggregating the RECS model’s predictions of AC access to the state level and the synthetic U.S. population model’s predictions of AC access to the census tract level. Third, we evaluated tract-level disparities by plotting the aggregated tract-level predictions of AC access against CDDs, population density, and demographic composition (defined by the proportion of the tract’s population that are both below the Federal Poverty Line (FPL = \$25,000) and non-White) using hexbin plots. And finally, we evaluated individual-level disparities by plotting the raw individual-level predictions of AC access by age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, and building age using violin plots. This gave us a detailed view of how and why different groups may experience varying levels of AC access within the U.S. population.

4. RESULTS.

4.1 Study population.

Table 12 provides an overview of the descriptive statistics for the synthetic U.S. population used in this study. This population comprises 330,587,563 synthetic agents that reflect the real U.S.

population with respect to age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel type. The average age of the population was 39 years, with an equal distribution between males and females (50% each). The majority of the population identified as White (66%), followed by Black (12%), Asian (6%), and Native (1%), while 15% identified as “Other” (including mixed races or other races not captured in the aforementioned groups) and 19% identified as Hispanic. The mean household income was \$105,671, with a large proportion of households earning over \$75,000, though a wide income distribution points to significant economic variability.

Housing data revealed that the majority of the population owned their homes (66%) while the rest rented (34%). The housing stock was also dominated by single family homes (69%), followed by apartments (25%) and mobile homes (6%), with an average housing occupancy of 2.61 people. The age of buildings spans various decades, with a relatively even distribution across decades, reflecting historical shifts in residential development.

Lastly, there was substantial variability in climate conditions across census tracts, where the mean CDDs across tracts was 1,636 degrees with a standard deviation of 1,130.

Table 12. Descriptive statistics of the synthetic U.S. population dataset.

Characteristic	N = 330,587,563 ¹
Age (years)	39 (23)
Sex	
Female	166,805,584 (50%)
Male	163,781,979 (50%)
Race	

Characteristic	N = 330,587,563¹
White	217,813,030 (66%)
Asian	19,070,789 (6%)
Black	41,053,109 (12%)
Native	3,376,045 (1%)
Other	49,274,590 (15%)
Ethnicity	
Hispanic	61,573,472 (19%)
Non-Hispanic	269,014,091 (81%)
Household Income (dollars)	105,671 (115,704)
Household Income	
< \$5000	9,289,926 (2.8%)
\$5000-7499	2,727,435 (0.8%)
\$7500-9999	3,103,650 (0.9%)
\$10000-12499	7,521,202 (2.3%)
\$12500-14999	5,519,102 (1.7%)
\$15000-19999	10,435,006 (3.2%)
\$20000-24999	12,360,779 (3.7%)
\$25000-29999	11,220,761 (3.4%)
\$30000-34999	12,929,225 (3.9%)
\$35000-39999	11,454,475 (3.5%)
\$40000-49999	23,570,938 (7.1%)
\$50000-59999	21,947,991 (6.6%)
\$60000-74999	30,967,191 (9.4%)
\$75000-99999	42,359,016 (13%)
\$100000-149999	57,097,066 (17%)
> \$150000	68,083,800 (21%)

Characteristic	N = 330,587,563 ¹
Housing Tenure	
Own	217,030,162 (66%)
Rent	113,557,401 (34%)
Housing Occupancy (people)	2.61 (1.62)
Building Type	
Apartment	82,217,358 (25%)
Mobile Home	18,654,135 (6%)
Single Family Home	229,716,070 (69%)
Building Age	
Before 1950	52,971,945 (16%)
1950-1959	33,249,208 (10%)
1960-1969	33,329,736 (10%)
1970-1979	47,899,418 (14%)
1980-1989	44,286,086 (13%)
1990-1999	42,881,310 (13%)
2000-2009	41,616,779 (13%)
2010 or later	34,353,081 (10%)
CDD (degrees)	1,636 (1,130)
Limited Residential AC Access (probability)	0.24 (0.17)

¹Mean (SD); n (%)

4.2 Key predictors of residential AC access.

We found that the mean predicted probability of limited residential AC access was 0.24, though individual probabilities varied widely ($SD = \pm 0.17$), ranging from 0.03 to 0.97 across the synthetic U.S. population (1st quartile = 0.11, median = 0.20, 3rd quartile = 0.32). Given this

variability, we used a feature importance matrix to determine which features most strongly influenced the model's predictions. Household income emerged as the most significant predictor (24.14%), followed by state (18.37%), CDD (15.69%), and building age (12.31%). Additional features contributing to limited AC access included age (8.48%), housing occupancy (6.57%), race (4.86%), sex (4.45%), housing tenure (2.07%), ethnicity (1.72%), and building type (1.32%). These percentages represent the relative frequency of each feature's use in decision-tree splits across all trees in the gradient boosting model. Other metrics included in the feature importance matrix—including gain, which measures the improvement in accuracy brought by a feature when it is used for splitting, and cover, which quantifies the proportion of samples influenced by a feature across all splits—are provided in **Table 1B in Appendix B**.

4.3. Assessing the spatial distribution of residential AC access.

To illustrate the critical differences between using individual-level microdata at coarse spatial resolutions and fine spatial resolutions, we produced two maps of the individual predicted probabilities of residential AC access: one based on RECS data (**Figure 14a**) and the other using our synthetic U.S. population data (**Figure 14b**). **Figure 14a** highlights that Western and Northwestern states, including California, Oregon, Washington, Idaho, Montana, and Wyoming, appear most vulnerable to limited residential AC access. However, **Figure 14b**, reveals that state-level averages not only mask important census-tract-level variability within these regions but also dilute perceptions of vulnerability across Southwestern, Southern, Midwestern, and Northeastern states. This map shows that even within the same county, differences in demographic composition, housing quality, resource availability, energy costs, and climate conditions can significantly influence residents' ability and need to access AC.

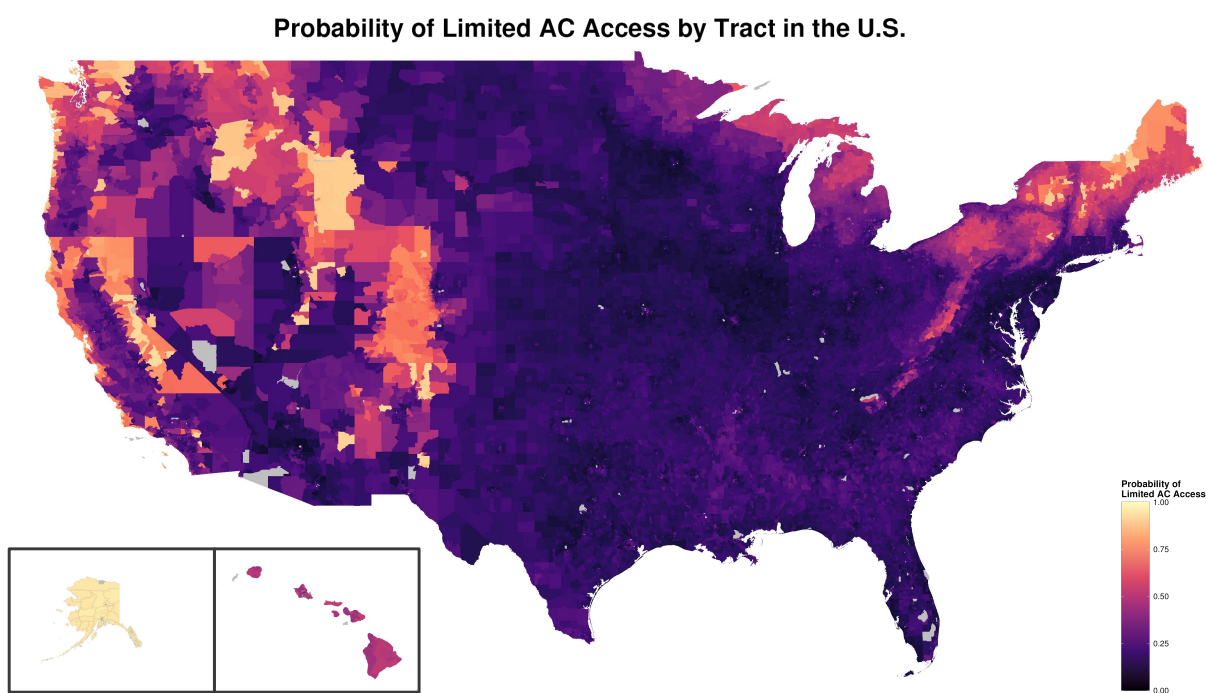
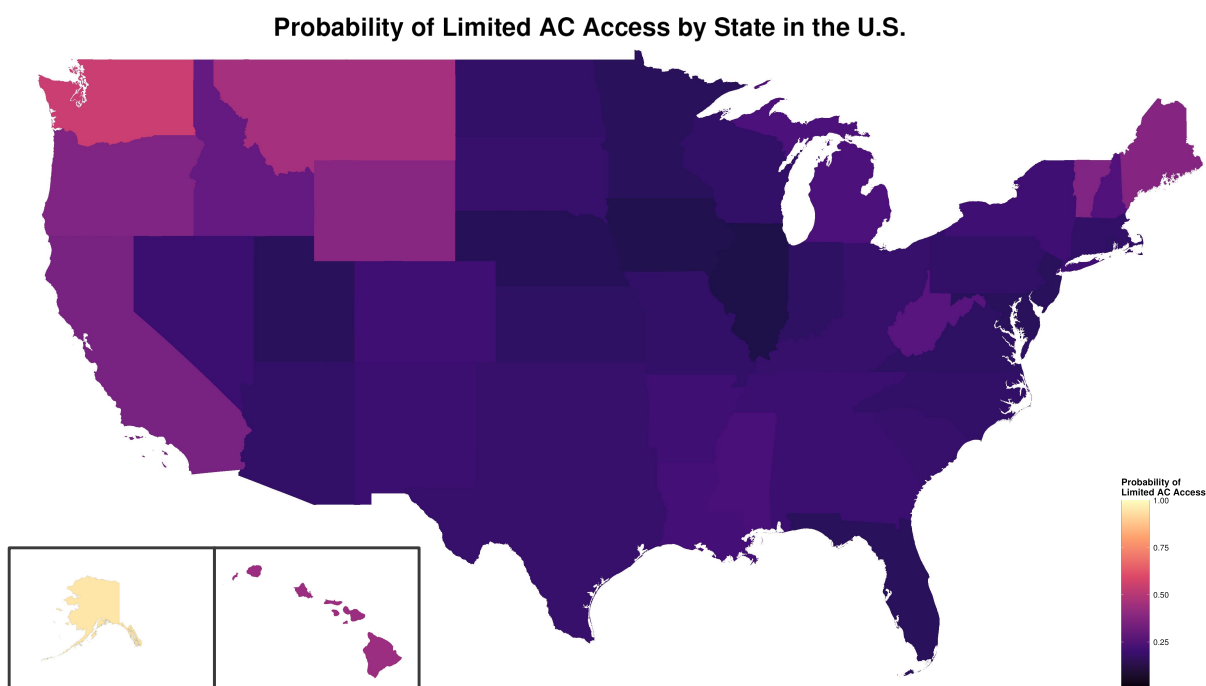


Figure 14. Predicted probability of limited residential AC access modeled as a joint function of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDDs, and state using (a) microdata drawn from RECS at the state level versus (b) microdata drawn from the synthetic U.S. population at the census tract level.

4.4. Assessing the influence of tract-level characteristics on residential AC access.

Next, we assessed the relationship between CDDs and access to AC at the census tract level. **Figure 15a** shows that most tracts are concentrated at lower CDDs (below 2,000 CDDs), where the probability of limited residential AC access often exceeds 0.25. The probability of limited residential AC access declines sharply as CDDs increase, clustering around 0.10 for tracts with CDDs over 2,000. This trend plateaus beyond this threshold, suggesting that tracts have adapted to meet cooling needs once a certain level of heat is reached.

Figure 15b illustrates the relationship between population density and access to AC. Most tracts are concentrated at lower population densities (below 1,000 people/km²), where probabilities of limited residential AC access are generally lower but vary widely from 0 to 1. As population density increases beyond approximately 25,000 people/km², the range of probabilities narrows and stabilizes within the 0.1 to 0.3 range. This pattern suggests that while lower-density tracts may show greater variability in access to AC, higher-density tracts tend to have probabilities of limited residential AC access that are generally lower and more consistent.

We then assessed the role of racial, ethnic, and economic composition in determining access to AC. **Figure 15c** shows that most tracts are concentrated at lower percentages of non-White, low-income residents (under 25%), where probabilities of limited residential AC access are typically lower but vary widely from 0 to 1. As the percentage of non-White, low-income residents increase, a more pronounced linear trend emerges, with probabilities of limited residential AC access consistently rising and clustering around 0.25. This trend indicates that as the proportion of racially- and ethnically- minoritized, economically-disadvantaged residents in a tract increases, the probability of limited residential AC access becomes not only higher but also more consistent.

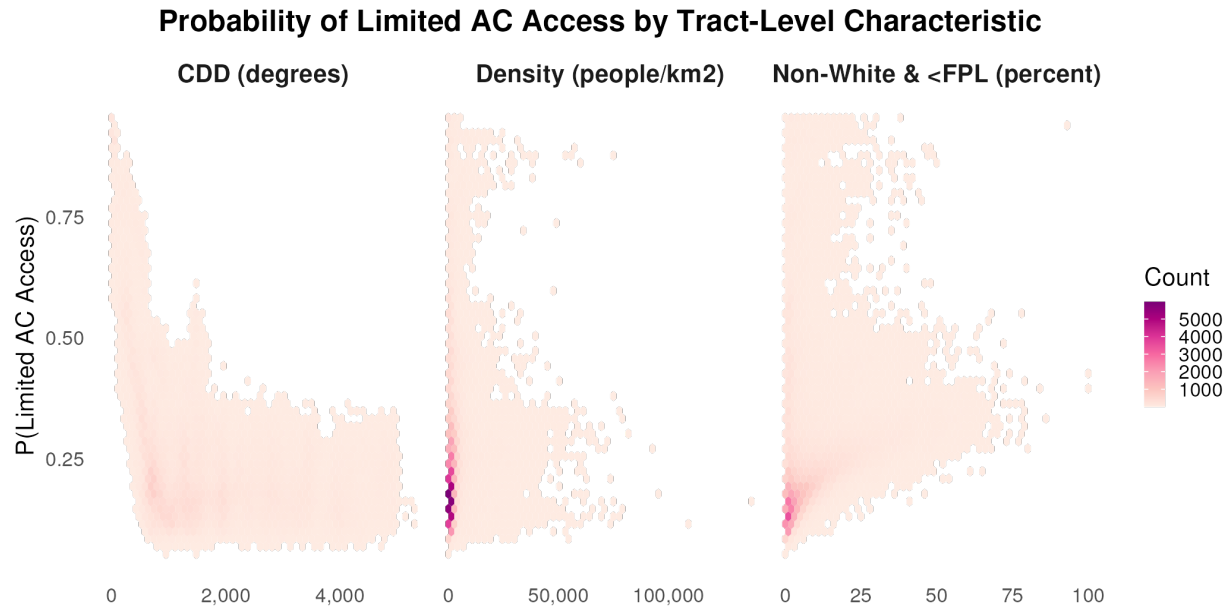


Figure 15. Predicted probability of limited residential AC access by census-tract-level characteristics, including (a) annual CDD (b) population density, and (c) the percentage of the population that is both non-White and below the federal poverty level.

4.5. Assessing the influence of individual-level characteristics on residential AC access.

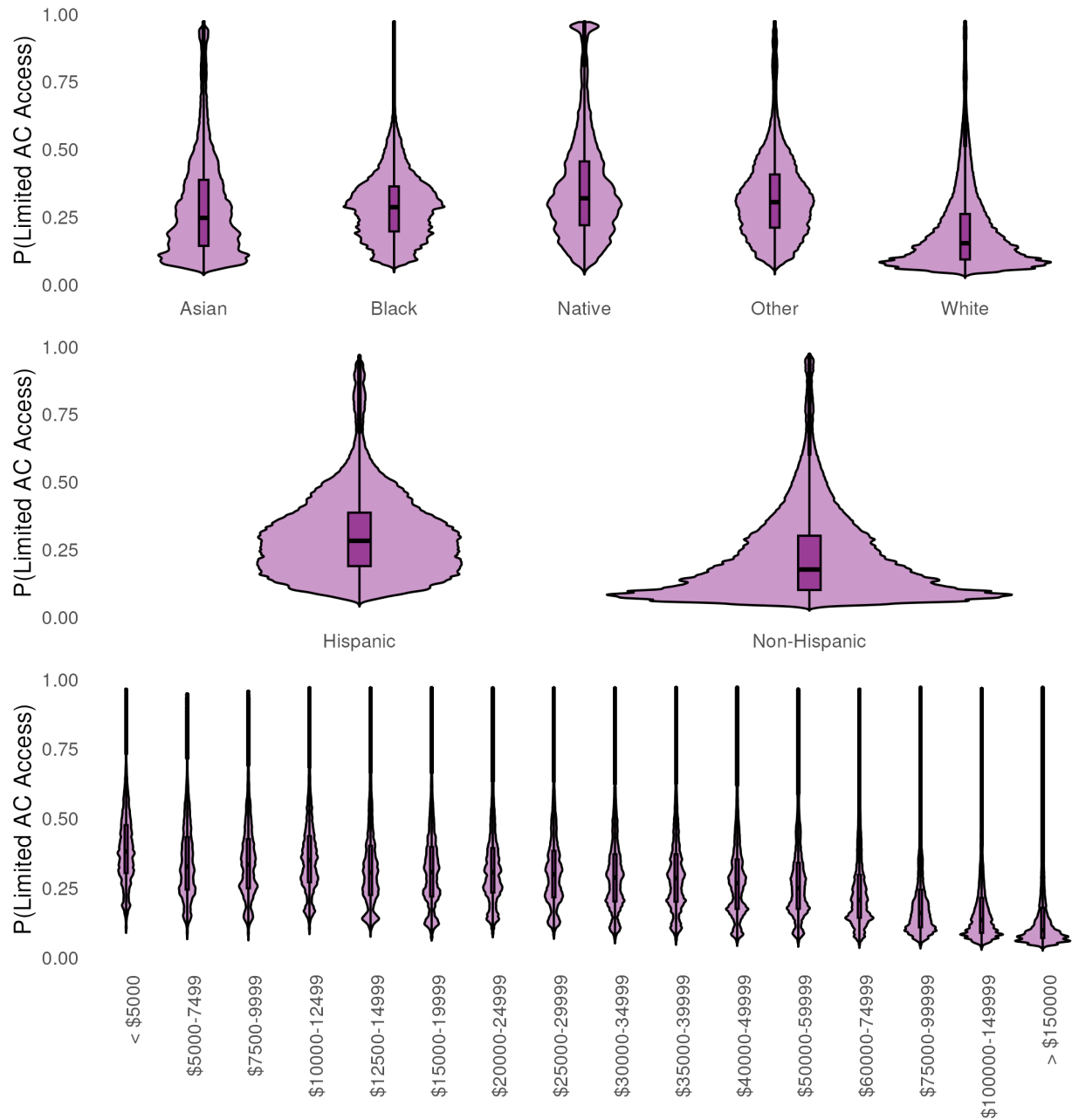
While there were no significant differences in mean probabilities of limited residential AC access by age or sex—children: 0.26 (SD ± 0.17), adults: 0.24 (SD ± 0.16), elderly: 0.20 (SD ± 0.17); female: 0.24 (SD ± 0.17), male: 0.24 (SD ± 0.17)—we observed substantial variability across race, ethnicity, and household income (**Figures 16a-16c**). Individuals identifying as Native had the highest mean probability of limited AC access (0.36), followed by Other (0.33), Black (0.29), Asian (0.29), and White (0.21), indicating a notably higher likelihood of limited AC access among non-White populations compared to White populations (**Figure 16a**). Similarly, Hispanic populations had a mean probability of 0.31, 8% higher than that observed for non-Hispanic populations (0.23) (**Figure 16b**). The violin plots show that probabilities for Asian, White, and non-Hispanic populations are more concentrated around lower values, whereas Native, Other, Black, and Hispanic

populations exhibit wider distributions. Notably, Native and Asian populations display a ‘bulb’ shape at the upper end of their distributions, signaling a subset with markedly high probabilities of limited AC access within these racialized groups.

A clear gradient in the probability of limited AC access across household income groups was also evident. The lowest household income group earning <\$5000 (0.40) had a mean probability 24% higher than that of the highest income group earning >\$150,000 (0.16) (**Figure 16c**). Additionally, as income increases, the probability distributions become more concentrated around lower values, suggesting that higher-income households experience not only lower, but also more consistent probabilities of limited residential AC access compared to lower-income households.

Housing characteristics also revealed meaningful differences in the probability of limited residential AC access (**Figures 16d-16g**). Renters were more likely to lack access compared to homeowners, with mean probabilities of 0.30 and 0.21, respectively (**Figure 16d**). This trend extended to both building type and age. Those living in apartments (0.29) and mobile homes (0.27) had a higher mean probability of limited residential AC access compared to those in single-family homes (0.22) (**Figure 16e**). A clear gradient was observed for building age, where homes built before 1950 showed a mean probability of 0.29, which steadily decreased with each subsequent construction era, reaching a mean probability of 0.16 for buildings constructed in 2010 or later (**Figure 16f**). Housing occupancy also impacted access, with probabilities increasing from 0.23 for households with one to two people to 0.30 for households with five or more people (**Figure 16g**). In general, subcategories with less access to AC—such as renters, apartment dwellers, and those residing in older or more crowded homes—exhibited greater variability in their distributions compared to those with better residential AC access, mirroring the trends observed in social and economic characteristics. Predicted probabilities across all characteristics can be found in **Table 2B in Appendix B**.

Probability of Limited AC Access by Individual-Level Characteristic



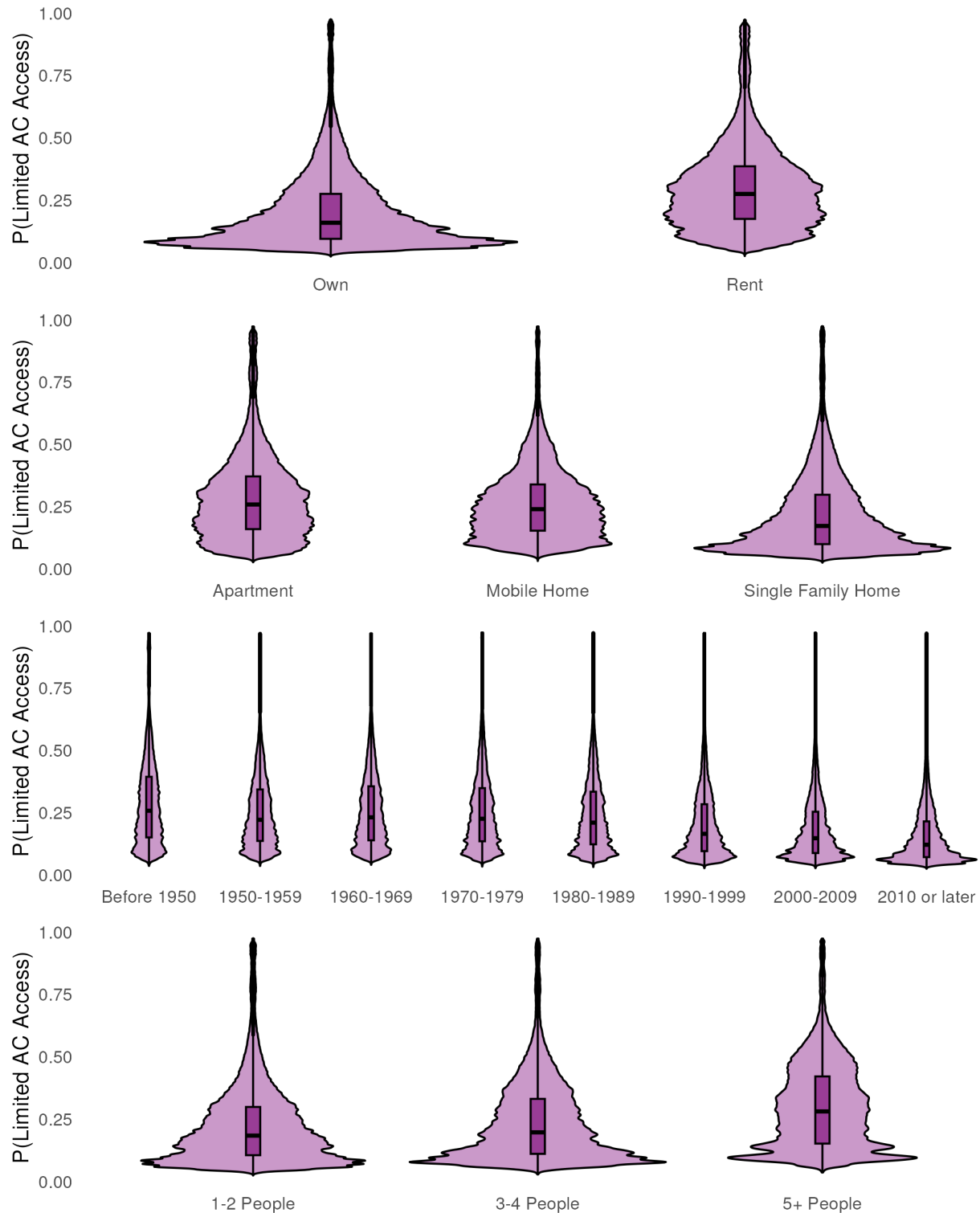


Figure 16. Predicted probability of limited residential AC access by individual-level characteristics, including (a) race (b) ethnicity, (c) household income, (d) housing tenure, (e) building type, (f) building age, (g) housing occupancy.

5. DISCUSSION.

To our knowledge, this study is the first to map individual probabilities of residential AC access—a comprehensive measure that accounts for one’s ability to own, install, maintain, and operate their AC system—across all U.S. census tracts. We achieved this by leveraging a combination of microdata from RECS and our synthetic U.S. population. Using machine learning techniques, we modeled residential AC access as a joint function of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDDs, and state, achieving strong predictive accuracy that allowed us to downscale individual residential AC access data that traditionally only provides spatial information at the state level to the census tract level. Our findings show that the average probability of limited residential AC access across the 330 million synthetic agents in the synthetic U.S. population was 0.24, with probabilities varying substantially (0.03–0.97), driven by social, economic, infrastructural, and environmental disparities. In addition to household income, building age, and CDDs, state emerged as a significant contributor to the model’s predictive power, likely reflecting state-level variations in cooling-related costs, policies, programs, practices, building standards, and energy infrastructures. Importantly, however, we revealed that state-level data often masks the true landscape of residential AC access, with our tract-level predictions uncovering substantial interstate and intrastate variability across all regions. For instance, while lack of residential AC access was most concentrated in the West, Northwest, and Northeast, significant disparities persisted within these regions in addition to the South, Southwest, and Midwest, underscoring the need to consider finer geographic scales when assessing AC access. This was driven by clear disparities in the probability of limited residential AC access based on climate conditions, population density, and demographic composition at the tract level alongside socioeconomic vulnerabilities at the individual level. These findings offer nuanced insights into the complex interplay of social, economic, infrastructural, and environmental characteristics that shape

inequities in residential AC access at fine spatial resolutions across the U.S. In this section, we situate these findings within the broader literature and illustrate their practical implications through case studies of three states, including Washington (WA), New York (NY), and Texas (TX).

To date, a several studies have sought to model residential AC access at fine spatial resolutions to better identify heat-vulnerable populations, but these studies face significant limitations that may affect their accuracy and applicability. First, existing research is often forced to rely on AC ownership as a stand-in for AC access, which may bias perceptions of how widespread AC access is. For example, an analysis by Romitti et al. (2022) on residential AC ownership in 115 metropolitan areas across U.S. census tracts reported a median predicted probability of residential AC ownership of 0.97, with a range from 0.15 to 1.00,²⁹¹ which is consistent with other research showing similarly high levels of residential AC ownership across the U.S. that exceed 0.90.^{301,302,305–308} In contrast, our study found a lower median predicted probability of residential AC access of 0.80, with a broader range from 0.03 to 0.97. This 17% decrease in the median and an 9% expansion in the range can primarily be attributed to differences in how the target variable was measured in existing studies compared to our study. Specifically, by relying on residential AC ownership as a proxy for residential AC access, existing studies may overlook the financial and logistical realities that affect whether households can actually use their AC systems. In contrast, by modeling whether households can afford to own, install, maintain, and operate their AC systems based on their joint social, economic, infrastructural, and environmental characteristics, we move beyond these simple proxies to capture the complexities that drive residential AC access in different subpopulations. Our approach therefore avoids overestimating residential AC access and provides a more comprehensive understanding of the variability in this access across the U.S.

Second, residential AC ownership data is typically restricted to either individual-level microdata at coarse spatial resolutions or geographically-aggregated data at finer spatial resolutions

to protect respondent privacy. Individual-level microdata provide joint distributions that reveal how multiple characteristics interact within individuals, whereas geographically-aggregated data provide marginal distributions that represent the average prevalence of single characteristics within an area (e.g., the percentage of residents within a census tract in a specific income bracket or in a specific building type), without showing how these characteristics interact within that area (e.g., the percentage of residents within a census tract in a specific income bracket *who live in a* specific building type). Romitti et al. (2022) made strides in addressing these data limitations by integrating both individual-level microdata and geographically-aggregated data in their analysis.²⁹¹ They utilized American Housing Survey (AHS) microdata containing joint distributions of multiple characteristics to model the joint effect of age, race, ethnicity, educational attainment, household income, housing tenure, building type, building age, and market value on individual probabilities of AC ownership at the Core-Based Statistical Area (CBSA) level.²⁹¹ This model was then applied to ACS data containing marginal distributions of these same characteristics to estimate average probabilities of AC ownership at the census tract level.²⁹¹ However, applying a model developed from joint distributions at the CBSA level to marginal distributions at the census tract level assumes that the relationships observed between characteristics at larger geographic levels also hold at smaller geographic levels. This assumption overlooks local variations in social, economic, infrastructural, and environmental contexts that can alter how these characteristics interact. A similar limitation exists in the development of heat vulnerability indices, where researchers typically use ACS data to overlay marginal distributions of “risk factors”—such as AC ownership, income levels, building types, heat exposures—at the census tract level to identify heat-vulnerability hotspots.^{301,302} This approach also assumes that the relationships observed at the area level apply to individuals within those areas. In both cases, the ecological fallacy arises when aggregate-level patterns are assumed to reflect individual-level interactions accurately, potentially masking or distorting the complex interactions

that shape patterns of heat vulnerability across U.S. subpopulations. Our approach counters this by applying a model developed from individual-level joint distributions with spatial information at the state level to individual-level joint distributions with spatial information at the census tract level, retaining the interactions between social, economic, infrastructural, and environmental characteristics at both geographic levels. This enables a more precise identification of heat-vulnerable populations, supporting the development and deployment of targeted interventions and resources.

While limited availability of replication data makes direct comparisons with our findings challenging—as existing studies have only assessed AC ownership in specific geographic areas within the U.S. following different methodological approaches—the findings of existing studies still provide useful information for contextualizing our own.

We first explored how disparities in residential AC access manifest at different spatial scales. Although existing studies may provide important localized insights by focusing on AC ownership in areas such as Detroit,³⁰⁵ San Diego County,³⁰⁶ Los Angeles County,³⁰⁷ Maricopa County,³⁰⁸ California,³⁰³ and Florida³⁰⁴, their findings may not generalize well to other areas. This may be particularly true in other states, if we consider our findings on the significant contribution of the state variable to our model’s predictive power (18%). This suggests that unobserved state-level factors—including state-level differences in cooling-related costs, policies, programs, practices, building standards, and energy infrastructures—likely play a crucial role in shaping residential AC access. Romitti et al. (2022) took a broader geographical approach by producing AC ownership predictions in urban areas across states, identifying a latitudinal trend where AC ownership was highest in the South and lower in other regions,²⁹¹ which aligns with our findings that census tracts with lower CDDs tend to have reduced residential AC access. This can be attributed to the Southern states’ long history of adapting to high temperatures through widespread adoption of central AC as a

standard feature in homes and commercial buildings. In contrast, Western, Northwestern, and Northeastern states which have historically experienced milder climates, are now struggling to integrate effective cooling solutions as temperatures rise, often relying on window AC units that are both less efficient and less effective in maintaining safe indoor temperatures. However, our study extends this understanding by expanding AC access predictions beyond urban centers to also include rural areas, where 33% of the U.S. population resides. This broader coverage is essential, as our findings reveal that census tracts with lower population densities often face reduced residential AC access. This disparity may stem from economic constraints, fewer available AC service providers, and limited energy infrastructure characteristic of rural areas.³²¹ Together, these findings highlight the strengths of our approach in capturing disparities in residential AC access both between and within states, offering a comprehensive view of how these disparities unfold at fine spatial resolutions across the entire U.S.

We then explored how disparities in residential AC access manifest for different subpopulations experiencing compounding social, economic, infrastructural, and environmental vulnerabilities. Census tracts with higher percentages of residents who both identified as racial or ethnic minorities and lived below the FPL demonstrated a strong positive relationship with lack of access to residential AC. This finding set the stage for examining the specific individual characteristics that contribute to these disparities in residential AC access. Household income emerged as the most influential predictor in our model, contributing 24% to its predictive power and revealing a pronounced gradient, where lower-income households had a significantly higher probability of limited residential AC access compared to higher-income households. For example, the lowest income group was 24% more likely to experience limited residential AC access than the highest income group. Racial and ethnic disparities were also evident, with individuals identifying as Native, Black, Asian, Other, and Hispanic consistently exhibiting mean probabilities at least 9%

higher than their White and non-Hispanic counterparts. Housing-related vulnerabilities compounded these disparities. Building age, contributing 12% to the model's predictive power, showed a clear trend where those living in older buildings were up to 13% more likely to experience limited residential AC access compared to those living in newer buildings. Other characteristics typical of those facing housing-related vulnerabilities—including renter status and apartment dwelling—exhibited both higher probabilities of limited residential AC access and greater variabilities in their distributions compared to their respective counterparts. These issues may reflect difficulties in improving residential AC access in older buildings, especially for renters and residents of multi-unit dwellings, who often lack the authority or resources to make structural changes.^{322–324} While this suite of findings reinforces the understanding that marginalized populations tend to have limited access to health-promoting resources,^{278,279,289–299} our approach uniquely captures how residential AC access is produced by compounding social, economic, infrastructural, and environmental vulnerabilities, allowing us pinpoint which groups are most vulnerable and where they reside.

Our findings, which reveal substantial disparities in residential AC access driven by social, economic, infrastructural, and environmental characteristics, highlight the pressing need for targeted policy responses tailored to diverse community needs across the U.S. Translating these insights into practical action is essential for optimizing resource allocation and enhancing public health interventions. Existing programs, such as the Department of Energy's (DOE) Weatherization Assistance Program (WAP) and the Department of Health and Human Services' (HHS) Low Income Heat Energy Assistance Program (LIHEAP), lay a foundation for expanding residential AC access by supporting costs associated with AC use but face limitations that hinder their reach and impact.³²⁵ WAP plays a critical role in addressing cooling needs beyond AC ownership, providing assistance for AC system repairs, insulation upgrades, and window replacements³²⁶, all of which are

essential services to maintaining safe indoor temperatures. Yet, WAP can serve only 35,000 homes annually at current funding levels,³²⁶ leaving many vulnerable households without support. Our model, which can pinpoint vulnerable populations struggling specifically with AC usage, could enable WAP to maximize its impact within a constrained budget by prioritizing weatherization assistance for those most in need. LIHEAP, which has a broader reach, operates as a targeted block grant where the federal government distributes resources to state governments, who are responsible for dispersing funds directly to households or utility companies on the households' behalf to pay down home heating or cooling bills that they are unable to otherwise afford.³²⁷ In 2023, LIHEAP served 7.1 million American households, but this figure represents only 20% of eligible households needing cooling assistance, leaving a funding shortfall of over \$2 billion.³²⁸ Moreover, only 24 states currently offer cooling assistance, which accounts for just 9.8% of the overall LIHEAP budget³²⁹—a reflection of the program's historical emphasis on heating needs. This imbalance can be attributed to the outdated formulas governing state distributions.²⁷⁷ While an updated formula established in 1984 aimed to balance heating and cooling needs by incorporating both heating degree days (HDDs) and CDDs, longstanding “hold-harmless” provisions have maintained funding levels based on an outdated formula focused on heating needs from 1981, and this is further entrenched by Congressional decisions to override the new formula in an effort to preserve funding in cold-weather states.²⁷⁷ These provisions result in funding allocations that are misaligned with current energy burdens, leaving many populations underserved as the duration of the hot season and the frequency and severity of heat waves continue to increase.²⁷⁷ Addressing this misalignment calls for more granular data on AC access to guide states in efficiently allocating limited resources and effectively developing tailored interventions for the most heat-vulnerable populations. We illustrate the practical implications of using our model to generate such data through case studies of WA, NY,

and TX, demonstrating opportunities for it to strengthen programs like LIHEAP and WAP in diverse contexts.

Washington.

While state-level data indicates that much of Washington's population faces challenges in accessing AC, our analysis at the census tract level reveals significant regional variability. Urban areas in Western Washington show markedly higher probabilities of limited residential AC access compared to areas in Eastern Washington (**Figure 17**). This disparity is particularly concerning given the 2021 Pacific Northwest heatwave. The historically mild climate of Western Washington has meant that many buildings were constructed without AC. During the heatwave, disproportionately severe conditions³³⁰ coupled with the lack of protection³³¹ resulted in hundreds of deaths in urban areas across Western Washington like Seattle and King County, particularly affecting marginalized populations.^{332,333} Since then, Washington has prohibited utility shutoffs above 90°F³³⁴ and updated building codes to require heat pumps and set a maximum indoor temperature of 78°F for new buildings.³³⁵ However, Washington does not receive LIHEAP cooling assistance given that it was historically a cold-weather state,³³⁶ leaving low-income residents in older buildings outside the scope of these new standards. Our tool can help guide LIHEAP cooling assistance, retrofit incentives, AC deployment, shading initiatives, and community cooling center placement to protect these populations as the frequency and severity of heatwaves increase across the Northwest.

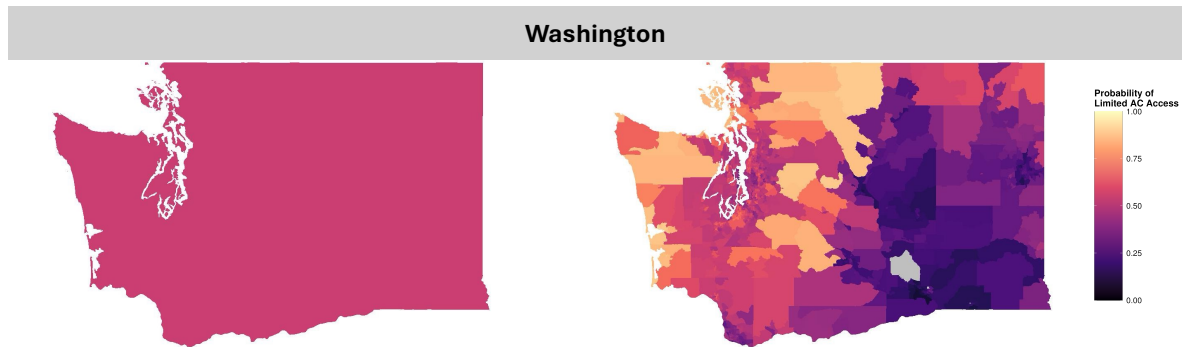


Figure 17. Predicted probability of limited residential AC access in WA using (a) microdata drawn from RECS at the state level versus (b) microdata drawn from the synthetic U.S. population at the census tract level.

New York.

In New York, high-density urban areas like New York City (NYC), and to a lesser extent, those along the I-90 corridor from Albany to Buffalo generally experience high levels of AC access, which can obscure significant disparities that persist in lower-density rural areas when looking solely at a state-level map compared to a tract-level map (**Figure 18**). Despite this perception of widespread AC access, New York has the nation’s highest concentration of older buildings,³³⁷ many of which lack modern design features and central cooling systems essential for maintaining safe indoor temperatures.³³⁸ As a result, residents in both urban and rural areas often rely on window AC units that are less effective and more costly, contributing to approximately 350 heat-related deaths annually.³³⁹ Cooling assistance is also limited, with only 2% of New York’s LIHEAP funding allocated to cooling needs.³³⁶ A new Tenants’ Right to Cooling bill introduced in 2024 aims to address this gap by requiring landlords to maintain indoor temperatures below 78°F when it exceeds 82°F outside, with a four-year compliance window for structural upgrades.³⁴⁰ Until these standards take effect, NYC’s Home Energy Assistance Program (HEAP) continues to provide free AC

units to eligible residents in need.³⁴¹ Our tool can further support these efforts by identifying areas where socioeconomic and housing vulnerabilities intersect, enabling targeted interventions tailored to specific contexts. In urban areas, it can help expand NYC HEAP's AC distribution programs to high-risk neighborhoods, while in rural regions, it can guide LIHEAP cooling assistance, to better protect residents as projected increases in temperature and humidity amplify heat stress across the Northeast.

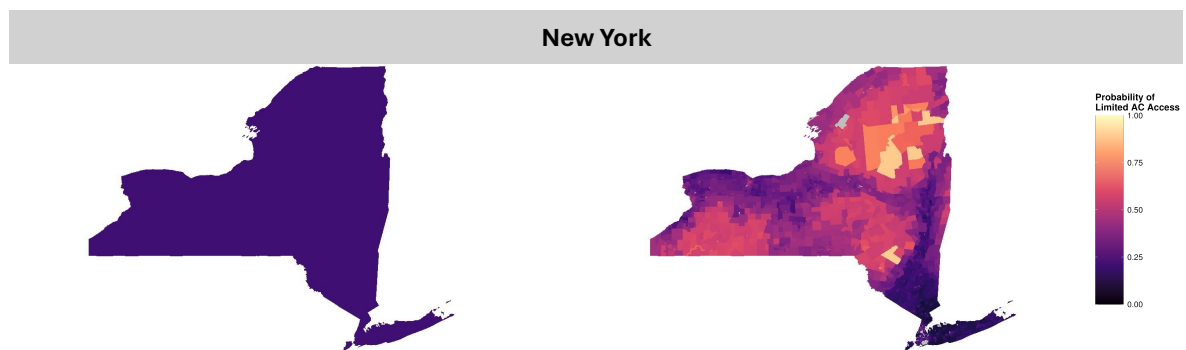


Figure 18. Predicted probability of limited residential AC access in NY using (a) microdata drawn from RECS at the state level versus (b) microdata drawn from the synthetic U.S. population at the census tract level.

Texas.

Texas presents an interesting case in our analysis of residential AC access disparities. The state- and tract-level maps reveal minimal differences indicating generally strong residential AC access across most regions due to historically high temperatures and the widespread adoption of cooling infrastructure (**Figure 19**). However, this does not mean that Texas should be overlooked. Our model highlights notable disparities along the southern border, where tracts with higher probabilities of limited residential AC access coincide with communities that are predominantly low-income and Hispanic or Latino. These communities often suffer from limited economic opportunities, restricted access to

assistance programs, and severely under-resourced infrastructures, making them particularly vulnerable to heat stress and heat-related illness.³⁴² Coupled with the fragility of Texas’s energy grid—which has faced significant strain and failures during extreme weather events like the 2021 winter storm and an increasing number of heatwaves—the projected increase in heat stress across the South is expected to exacerbate existing vulnerabilities.^{343–345} Addressing these needs involves not only strengthening the cooling infrastructure but also fortifying the energy grid to support sustained AC use during critical periods. Our tool can guide these efforts by pinpointing tracts that, while appearing adequately equipped at a state level, face distinct challenges that require proactive solutions such as the implementation of community cooling centers and targeted cooling assistance.²⁷⁷

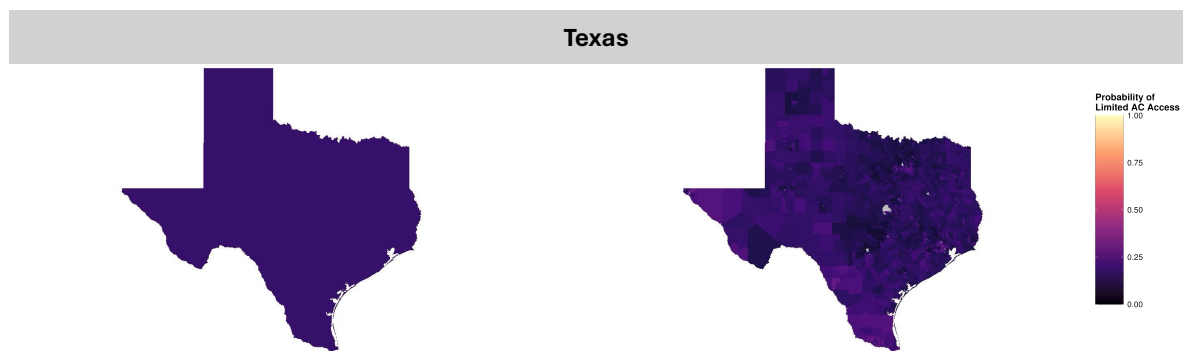


Figure 19. Predicted probability of limited residential AC access in TX using (a) microdata drawn from RECS at the state level versus (b) microdata drawn from the synthetic U.S. population at the census tract level.

5.1. Strengths.

We make substantial progress in understanding and mapping residential AC access across the entire U.S. through the integration of high-quality data sources and advanced modeling techniques. Unlike prior research that relied on residential AC ownership as a proxy for residential

AC access, our work provides a more comprehensive measure of true residential AC access, defined by the ability to own, install, maintain, and operate AC systems. We achieved this by leveraging microdata containing this target variable from RECS alongside microdata from our synthetic U.S. population, which simulates real-world distributions of social, economic, infrastructural, and environmental characteristics at the census tract level. Using a gradient boosting model, optimized through extensive hyperparameter tuning to achieve high predictive accuracy (ROC-AUC of 0.95 for training and 0.80 for testing), we captured complex, non-linear interactions among key predictors such as age, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, CDD, and state. This approach bypassed the common assumption that relationships between these variables are linear or straightforward, resulting in a realistic representation of their joint effects on residential AC access at a fine spatial resolution. The model's generation of individual-level probabilities of residential AC access, covering the entire U.S. at the census tract level, provides an unprecedented granularity that reveals significant variations within and between tracts. This fine-grained distribution underscores how the social, economic, infrastructural, and environmental characteristics which underlie population health dynamics intersect to influence residential AC access. These insights are invaluable for policymakers, public health officials, and urban planners, equipping them with the precision needed to identify heat-vulnerable populations.

5.2. Limitations.

This study contributes valuable insights into disparities in residential AC access across the U.S.; however, several limitations must be acknowledged. First, the analysis depends on the quality and scope of the underlying datasets, each of which has inherent constraints. The RECS dataset, although instrumental in providing household energy data, is based on self-reported survey

responses from a relatively limited sample of the U.S. population. This reliance may introduce biases, particularly if the sample does not fully represent broader population characteristics. Additionally, while the synthetic U.S. population has been validated to reflect key demographic, socioeconomic, and environmental distributions in the real U.S. population at the census tract level (as detailed in Parikh et al., 2024), it may still carry biases inherent to its generation process that could impact the accuracy of model predictions. Second, the integration of household-level RECS data with individual-level synthetic U.S. population data required several assumptions. Specifically, race and ethnicity were assumed to be consistent across household members based on data showing that over 90% of U.S. households are racially or ethnically homogeneous, while sex was probabilistically assigned using state-level distributions from the ACS. While these assumptions facilitated data integration, they may overlook nuanced intra-household diversity, potentially impacting the precision of AC access predictions. Third, the target variable “no AC days” may be subject to recall bias, though it’s unclear if there is a directional bias. To address data imbalances and improve model performance, we transformed this variable into a binary measure, which captures whether households had access to AC but does not reflect the full range of deprivation. Fourth, while we included a broad range of social, economic, infrastructural, and environmental characteristics in our model, some important characteristics influencing residential AC access were only partially or not at all captured. This limitation stemmed both from the availability of variables and the need to use variables common to both data sources. For example, we used state as a proxy to capture variations in cooling-related costs, policies, programs, practices, building standards, and energy infrastructures, but could not disentangle which of these state-level characteristics most directly impacted residential AC access. We also used CDDs as a measure of temperature exposure but recognize that this metric does not fully represent environmental conditions prompting AC use, as it excludes humidity levels which can significantly affect heat perception. Additionally, income

does not fully capture wealth, an important distinction when evaluating economic barriers to AC access. Finally, we could not include structural characteristics such as building size, insulation quality, reflective coatings, and shading mechanisms, as these were only available in the RECS dataset, yet they play a critical role in residential AC use. The absence of such variables may have affected the precision of our model's predictions and addressing these gaps represents a critical area for future research. Fifth, while residential AC access is associated with reduced heat-related mortality, this study does not capture all characteristics that contribute to heat stress exposures or heat stress resilience. By focusing solely on residential AC access, other critical elements such as access to workplace AC systems, community cooling centers, passive cooling solutions, adaptive behaviors, and biological susceptibility were not modeled, limiting the scope of our findings to disparities in residential AC access alone. Sixth, while gradient boosting was chosen for its strong ability to model complex, non-linear interactions and achieve high predictive accuracy (training ROC-AUC of 0.95 and testing ROC-AUC of 0.80), it comes with limitations. Its "black box" nature means that, although it effectively captures intricate relationships among variables, it offers limited interpretability, making it difficult to discern the relative contributions or specific exposure-response functions of individual predictors. This highlights the trade-off between predictive performance and transparency when applying machine learning techniques in public health research. Seventh, this study provides a snapshot based on the most recent RECS and synthetic U.S. population datasets. However, as demographic, socioeconomic, and climate conditions evolve, the landscape of residential AC access may shift significantly. Future work should aim to adapt the model to reflect these changing conditions, enabling more forward-looking analyses that can guide interventions as new vulnerabilities emerge.

Lastly, despite the significant benefits of AC, the widespread and growing reliance on AC amid prolonged hot seasons and more frequent and severe heatwaves creates a reinforcing cycle

with significant repercussions that must be considered. Increased AC use drives up electricity demand, placing strain on the electric grid.^{277,280,281,346–348} This strain has contributed to a tenfold increase in power outages and rolling blackouts—situations that are particularly dangerous during extreme heatwaves.³⁴⁶ If the additional power required to alleviate this strain is sourced from fossil fuels, greenhouse gas emissions will rise, and so too will anthropogenic heat.²⁸⁰ Yet, while alternative strategies for cooling have lower greenhouse gas emissions than AC, they may not be sufficient to offset the physiological strain of heat stress amongst vulnerable groups including children, the elderly, and those with pre-existing health conditions.²⁹¹ Therefore, while our study underscores the critical role of residential AC access in protecting against heat stress, it also highlights the need for strategies that enhance equitable access to cooling while addressing the public and environmental health impacts of rising AC use.³⁴⁸

6. CONCLUSIONS.

Residential air conditioning (AC) is an important tool for protecting against the harmful effects of heat stress on population health and economic outcomes. Our study provides a pioneering examination of disparities in residential AC access at a fine spatial resolution across the entire U.S., leveraging machine learning techniques applied to a combination of microdata from RECS and our novel synthetic U.S. population to assign individual probabilities of residential AC access to 330 million synthetic agents.

Our findings highlight that commonly used estimates relying on residential AC ownership as a proxy for residential AC access may significantly underestimate the scope of the problem, failing to capture the barriers to installation, maintenance, and operation that impact the ability to use residential AC. Furthermore, we observed significant variability in residential AC access at the state,

tract, and individual levels, driven by compounding social, economic, infrastructural, and environmental vulnerabilities. These disparities underscore the urgent need for targeted interventions and policies that consider the multifaceted nature of residential AC access to ensure equitable protection against heat stress across diverse U.S. communities.

Given the issues posed by increasing AC use across the U.S., addressing these challenges requires a multilevel, multipronged approach. Policy efforts should prioritize investment in alternative cooling strategies, including the enhancement of building insulation, adoption of reflective coatings, expansion of greenspace, and addition of shading, alongside the development of energy-efficient AC technologies. Simultaneously, the resilience of the energy grid must be bolstered to accommodate increased demand, with a parallel shift towards renewable energy sources to mitigate the environmental impact. The granular insights generated by our model can guide these initiatives, helping identify and prioritize the distribution of resources and interventions to communities in greatest need. This approach can foster climate resilience and promote climate equity, ensuring that all populations are better protected as heat stress becomes more common and severe in a changing climate.

CHAPTER 5: CONCLUSIONS

This dissertation set out to overcome the limitations of existing data by constructing, validating, and applying a novel synthetic U.S. population capable of uncovering compounding social, economic, infrastructural, and environmental vulnerabilities at the local level. The findings of this dissertation underscore the potential of this synthetic U.S. population to serve as a tool for researchers, policymakers, and practitioners to devise and disperse more effective and equitable public health interventions.

Summary of findings.

Study 1 focused on the development and validation of a novel synthetic U.S. population that reflects key social, economic, infrastructural, and environmental characteristics of the real U.S. population at the census tract level. Using IPF, we constructed a novel synthetic U.S. population of approximately 330 million agents, constraining the joint distributions of age, sex, race, ethnicity, household income, housing tenure, housing occupancy, building type, building age, and heating fuel type derived from the 2022 ACS PUMS to their respective marginal distributions derived from the 2018–2022 ACS at the census tract level. Validation metrics confirmed its strong performance, with Pearson’s R^2 averaging 0.98, NRSME averaging 0.11, SAE averaging 0.03 across census tracts in internal validation, and KL divergence statistics remaining below 0.08 across all variables in external validation. These findings demonstrate that our synthetic U.S. population can provide a robust foundation for spatially explicit analyses and serves as the basis for the subsequent studies.

Study 2 applied the synthetic U.S. population to identify children under the age of five who may be at increased risk of lead-based paint exposure from living in homes built before 1980—a well-established public health concern in the U.S. By leveraging the joint distributions inherent in

our synthetic U.S. population, we identified up to 73% more at-risk census tracts than traditional methods, representing an estimated 3.28 million additional at-risk children. These findings revealed both systemic inequities in risk detection, with traditional approaches disproportionately missing census tracts with higher concentrations of uninsured, White, and Native populations, and significant limitations of solely relying on housing age as a risk proxy for lead-based paint exposure without considering housing quality.

Study 3 extended the synthetic U.S. population to address disparities in residential air conditioning (AC) access—a critical tool for mitigating against the adverse effects of heat stress amidst rising global temperatures across the U.S. We integrated AC access data from the Residential Energy Consumption Survey (RECS) with our synthetic U.S. population using machine learning techniques, modeling individual probabilities of residential AC access, and effectively downscaling a target variable that was previously only available with spatial information at the state level to the census tract level. This analysis uncovered significant disparities tied to social, economic, infrastructural, and environmental characteristics at the state, tract, and individual levels, with marginalized populations facing lower and more variable probabilities of residential AC access than suggested by measures solely based on residential AC ownership.

Practical implications of findings

The three studies comprising this dissertation point to several practical applications of our novel synthetic U.S. population for public health research and policy. In Study 1, we constructed and validated a synthetic U.S. population to provide a foundational tool that offers three major capabilities to support public health research and practice. First, it enables the study of joint distributions between multiple characteristics at fine spatial resolutions, allowing for a deeper

understanding of compounding, localized risks that are typically obscured in traditional datasets. Second, it supports the integration of diverse external datasets, enhancing the versatility of the model in addressing a wide range of research questions. Third, it provides the flexibility to estimate target variables that are otherwise difficult to attain.

Through Studies 2 and 3, we demonstrate several practical ways in which we can leverage these capabilities to simultaneously help us understand compounding, localized risks. In Study 2, this tool identified census tracts with high concentrations of children under five living in homes built before 1980 and provided the flexibility to examine the specific characteristics of children (e.g., by race, ethnicity, household income, or insurance status) and their living conditions (e.g., building type, ownership status) affecting this risk. Compared to traditional approaches, this novel approach uncovered up to 73% more at-risk tracts and revealed that these omissions were systematic, driven by the limitations, assumptions, and biases of traditional data. These findings underscore the importance of specifically targeting overlooked groups, such as Native American and uninsured populations, who already face compounding risks. Government agencies are uniquely positioned to act on these insights. At the federal level, these findings can inform the Federal Lead Task Force's initiatives under the EPA's Federal Action Plan to Reduce Childhood Lead Exposures by improving the accurate identification and targeting of at-risk communities for BLL surveillance, primary care, outreach programs, and abatement services.²¹³ They can also support HUD in prioritizing funding for lead remediation for low-income households through the Lead-Based Paint Hazard Control and Lead Hazard Reduction grant programs, strengthening enforcement of the Lead Disclosure Rule in high-risk areas, and addressing overlapping housing risks that compound lead exposure risks in low-income household through the Healthy Homes Program.²¹⁵ At the state and local levels, these findings can provide a foundation for implementing lead-safe housing ordinances that require landlords to remediate lead hazards in rental units. They also point to opportunities to offer financial

assistance or tax incentives to low-income homeowners to remediate lead hazards in their own homes.

Study 3 extended this tool's capability to identify census tracts with high concentrations of individuals unable to afford cooling due to barriers in AC ownership, installation, maintenance, or operation. Unlike traditional approaches that can only assess interstate disparities in AC access, our novel approach uncovered substantial intrastate disparities in AC access across social, economic, infrastructural, and environmental characteristics, revealing specific localized risks of being unprotected against the adverse effects of heat stress. These insights can directly inform federal programs like LIHEAP³²⁷ and WAP³²⁶ in how they could most effectively disperse funds to states, and in turn, how states disperse funds to households in greatest need. For LIHEAP, these findings can help identify low-income households that experience prohibitively high energy costs, enabling targeted cooling subsidies or energy assistance during the hot season. Similarly, for WAP, these findings can help identify low-income households in need of retrofits to improve weatherization and energy efficiency. They also highlight the need for local governments to integrate targeted cooling strategies into their heat resilience planning initiatives. For example, counties could use these findings to identify census tracts where they could allocate funds to establish cooling centers, increase land cover, or provide subsidies for passive cooling technologies or energy-efficient AC units. These efforts could also extend to strengthening the electric grid's resilience to accommodate increased cooling demand, ensuring reliable energy access for vulnerable populations.

At a higher level, this work illustrates the versatility of our synthetic U.S. population as a tool for tackling a wide range of public health challenges beyond disparities in lead exposure and AC access. By realistically simulating real-world joint distributions at fine spatial resolutions, the synthetic U.S. population could be integrated into tools used to identify spatial vulnerabilities like the Center for Disease Control and Prevention (CDC) Social Vulnerability Index (SVI),⁷⁰ CDC

Environmental Justice Index (EJI),⁷¹ White House Climate and Economic Justice Screening Tool (CEJST),⁷² Environmental Protection Agency (EPA) EJScreen,⁷³ Federal Emergency Management Agency (FEMA) Resilience Analysis and Planning Tool (RAPT)⁷⁴, which are currently subject to the limitations of geographically-aggregated data. Beyond enhancing existing tools, our synthetic U.S. population opens new possibilities for researchers to analyze emerging public health challenges across environmental justice, climate justice, infectious disease control, chronic disease prevention, and future health projections. Furthermore, it offers policymakers a cost-effective tool to simulate the impacts of interventions before implementation, enabling data-driven decisions that maximize positive outcomes while minimizing unintended consequences.

Recommendations for future research

Looking ahead, research should focus on advancing the synthetic U.S. population's capabilities by refining its underlying methodologies. While it was constructed using “gold-standard” datasets, its continued relevance depends on integrating updated data releases from the U.S. Census Bureau, particularly as efforts are made to address the under-sampling of hard-to-reach populations that can introduce sampling biases and percolate through the modeling process. Additionally, while iterative proportional fitting (IPF) is a widely-favored, reliable, and efficient technique for spatial microsimulation, several emerging techniques for synthetic population generation may offer opportunities for improvement.⁵³ Multilevel IPF preserves hierarchical relationships across geographic or administrative scales, such as census blocks within census tracts, by incorporating constraints at multiple levels of aggregation. Similarly, IPU is designed to maintain household structures by simultaneously constraining individual- and household-level attributes. Together, these approaches could enhance the model's ability to reflect both micro-level relationships and macro-

level dependencies, increasing its flexibility for integration with data sources that operate at varying scales. Other techniques, such as P-MEDM and FBS, provide additional advancements in synthetic population generation. P-MEDM incorporates varying degrees of confidence into constraints, allowing for greater flexibility when working with uncertain or incomplete data. FBS employs an evolutionary algorithm to iteratively optimize the synthetic population's fit to observed distributions, offering solutions for complex datasets. While these methods show potential for improving the reliability, robustness, and flexibility of our synthetic U.S. population, they remain in early stages of development and require further validation before widespread adoption.⁵³ Finally, adapting the synthetic U.S. population from a static model to a dynamic model—such as agent-based or future population simulations—could further extend its utility. These advancements would allow public health researchers, policymakers, and practitioners to better understand and preempt evolving behaviors, conditions, and needs across the population.

Future research building on Study 2 should focus on expanding the synthetic U.S. population's ability to assess compounding risks associated with lead exposure. This study intentionally employed a simplified scenario to facilitate comparison between our novel approach and traditional methods; however, future work could incorporate additional characteristics available in the synthetic U.S. population—such as race, ethnicity, household income, insurance status, and housing conditions—when mapping residential lead exposure hotspots, validating these findings using real BLL data. Future research could also incorporate external data sources such as HUD's Deteriorating Paint Index,²⁵⁷ the NRDC's aging lead-based water pipes map,²⁵⁸ contaminated outdoor soil, and proximity to highways, industrial sites, mining operations, and waste zones to create a more comprehensive view of total lead exposure hotspots.

Likewise, future research building on Study 3 should focus on advancing the synthetic U.S. population's capacity to model localized risks associated with cooling access by incorporating

additional contextual factors influencing heat resilience. One priority is to refine the model's ability to disentangle the underlying causes of limited residential AC access, including ownership, installation, maintenance, and operational barriers, to better target interventions based on specific barriers experienced by specific subpopulations. Expanding the predictors used in the model to include additional environmental (e.g., land cover, altitude, wet bulb globe temperature), neighborhood (e.g., access to workplace, community center, or public space cooling resources), and infrastructural (e.g., building size, insulation quality, reflective coatings) characteristics could also improve its ability to more accurately and comprehensively predict heat vulnerability. However, incorporating these predictors into the model would require future iterations of the Residential Energy Consumption Survey (RECS) to include these data at a finer spatial resolution, as they are not available in the current iteration. In addition to refining the model, future work should explore how state-level variations in costs, policies, programs, practices, building standards, and energy infrastructures interact to shape heat vulnerability. Finally, as demographic, socioeconomic, and climate conditions evolve, future research should aim to adapt the model to reflect these changing conditions, enabling forward-looking analyses that can guide future interventions as new vulnerabilities emerge.

Our synthetic U.S. population provides a solid foundation for future work to test interventions aimed at addressing complex public health challenges. However, as Tanton (2017) so profoundly noted:

“Many problems seen in the world today require input from different disciplines, and have no one solution – there may be a range of solutions. These problems are sometimes called “wicked problems” (Head, 2008), and call for an integrated approach (Pearson, Norman, O’Brien, & Tanton, 2017). Examples of wicked problems are climate change, social injustice, entrenched disadvantage and healthcare. Wicked problems require social, economic and environmental approaches to be combined in order to provide information to policy makers on policies that may assist in resolving these problems, rather than offering isolated solutions to them... I would argue that the next range of spatial microsimulation models needs to start

linking different models together, so that impacts across different areas can be estimated. This work has started, [but] there is significant progress that needs to be done in this area.”⁵³

While this dissertation integrates social, economic, infrastructural, and environmental characteristics into a single synthetic U.S. population model, it is merely the first step toward future innovation. Next steps should involve linking this synthetic U.S. population to other simulation models to test how implementing specific interventions—such as funding for lead remediation, cooling subsidies, or infrastructure upgrades—would affect outcomes across health, environmental sustainability, economic resilience, and equity. By addressing “wicked problems” with equally “wicked solutions,” researchers and policymakers can better understand not only the direct impacts of interventions but also their cascading effects across interconnected ecosystems, supporting more effective and equitable public health interventions.

Concluding remarks

Data serves as the bedrock for identifying vulnerabilities, devising interventions, and dispersing resources in the field of public health. Yet, when data is fraught with limitations, assumptions, and biases—failing to capture the full complexity of social, economic, infrastructural, and environmental vulnerabilities at fine spatial scales—its flaws permeate through public health research and into public health practice, perpetuating the very inequities it seeks to address.

This dissertation illustrates the real-world consequences of these flaws: Study 2 demonstrates how traditional data not only fails to detect millions of children who may be at risk of residential lead-based paint exposure but also systematically overlooks already vulnerable subpopulations. Similarly, Study 3 demonstrates that traditional data on AC access both underestimates the scope of limited AC access and obscures critical disparities in limited AC access driven by social, economic, infrastructural, and environmental forces. In this way, the inequities embedded in data mirror the

inequities embedded in life, underscoring the urgent need for innovative tools to address these inequities.

This dissertation represents a significant step forward in bridging these gaps through the development, validation, and application of a novel synthetic U.S. population tailored to the field of public health. By capturing the intricate interplay of social, economic, infrastructural, and environmental characteristics at fine spatial scales, this tool equips researchers, policymakers, and practitioners with the means to uncover hidden vulnerabilities and design targeted solutions that meaningfully address the needs of those who need them most.

APPENDIX

Appendix A.

Table 1A. Results of measures quantifying the difference between tracts identified by the traditional marginal distribution approach and those identified by our novel joint distribution approach when “high-risk” is defined as if *either* marginal distribution was in at least the third quartile for the traditional marginal distribution approach.

IoU	% tracts identified as “high-risk” by our novel joint distribution approach but not by the traditional marginal distribution approach	% tracts identified as “high-risk” by the traditional marginal distribution approach but not by the our joint distribution approach
65.92%.	8.53%	29.76%

Table 2A. Results of measures quantifying the difference between tracts identified by the traditional marginal distribution approach and those identified by our novel joint distribution approach when “high-risk” is defined as if *both* marginal distributions were in at least the third quartile for the traditional marginal distribution approach.

IoU	% tracts identified as “high-risk” by our novel joint distribution approach but not by the traditional marginal distribution approach	% tracts identified as “high-risk” by the traditional marginal distribution approach but not by the our joint distribution approach
66.39%.	21.25%	19.12%

Table 3A. Results of the SAR model presented for the full population when “high-risk” is defined as if *either* marginal distribution was in at least the third quartile for the traditional marginal distribution approach.

Variable	Non-White	White
Intercept	1.968**** [1.878, 2.062]	1.765**** [1.671, 1.864]
% <100% FPL	0.821**** [0.766, 0.879]	0.812**** [0.760, 0.867]
% Uninsured	1.498**** [1.319, 1.701]	1.521**** [1.342, 1.724]
% White	—	1.126**** [1.091, 1.162]
% Black	0.895**** [0.864, 0.927]	—
% Asian	0.941 [0.867, 1.021]	—
% Native	1.312*** [1.116, 1.542]	—
% Hispanic	0.863**** [0.834, 0.894]	0.923**** [0.892, 0.955]
% Rent	0.768**** [0.727, 0.811]	0.779**** [0.737, 0.822]
% SFH	0.773**** [0.702, 0.766]	0.732**** [0.702, 0.764]

Table 4A. Results of the SAR model presented for the full population when “high-risk” is defined as if *both* marginal distributions were in at least the third quartile for the traditional marginal distribution approach.

Variable	Non-White	White
Intercept	1.629**** [1.564, 1.696]	1.586**** [1.513, 1.662]
% <100% FPL	0.835**** [0.787, 0.886]	0.851**** [0.804, 0.901]
% Uninsured	1.386**** [1.243, 1.546]	1.437**** [1.291, 1.600]
% White	—	1.019 [0.991, 1.046]

% Black	1.006 [0.976, 1.036]	—
% Asian	0.951 [0.887, 1.020]	—
% Native	1.360**** [1.183, 1.564]	—
% Hispanic	0.942**** [0.914, 0.971]	0.945**** [0.918, 0.974]
% Rent	0.803**** [0.766, 0.842]	0.813**** [0.775, 0.852]
% SFH	0.759**** [0.731, 0.788]	0.764**** [0.737, 0.793]

Appendix B.

Table 1B. Summary of feature importance matrix results aggregated by variable.

Variable	Gain	Cover	Frequency
Income	0.209	0.193	0.241
State	0.122	0.221	0.184
CDD	0.167	0.171	0.157
Building age	0.046	0.055	0.123
Age	0.158	0.107	0.085
Housing occupancy	0.049	0.067	0.066
Race	0.054	0.054	0.049
Sex	0.143	0.079	0.045
Housing tenure	0.034	0.032	0.021
Ethnicity	0.012	0.012	0.017
Building type	0.005	0.009	0.013

Table 2B. Summary of predicted probabilities of limited AC access by variable.

Variable	Category	Mean	Median	SD	IQR
Age	Adults	0.2430187	0.20384163	0.1623378	0.2000909
	Children	0.2648771	0.23099512	0.1730980	0.2537471
	Elderly	0.2014974	0.14407104	0.1678876	0.1606154
Sex	Female	0.2393954	0.19468641	0.1664788	0.2090843
	Male	0.2425494	0.20020616	0.1672392	0.2131436
Race	Asian	0.2935703	0.24730122	0.1961994	0.2439362
	Black	0.2919801	0.28692240	0.1269398	0.1664714
	Native	0.3649891	0.31995857	0.2038394	0.2360197
	Other	0.3264438	0.30536282	0.1605550	0.1965153
	White	0.2054736	0.15354103	0.1596642	0.1681073
Ethnicity	Hispanic	0.3063486	0.28305644	0.1575820	0.1974573

Variable	Category	Mean	Median	SD	IQR
Household income	Non-Hispanic	0.2259910	0.17704982	0.1653177	0.2005244
	< \$5000	0.3965887	0.38166630	0.1386350	0.1719675
	\$5000-7499	0.3495567	0.32769078	0.1419281	0.1887053
	\$7500-9999	0.3499715	0.33475637	0.1443545	0.1769816
	\$10000-12499	0.3648577	0.35084534	0.1415879	0.1655677
	\$12500-14999	0.3316534	0.30590373	0.1592419	0.1767518
	\$15000-19999	0.3214670	0.30786318	0.1505318	0.1782744
	\$20000-24999	0.3286145	0.30715537	0.1473526	0.1596754
	\$25000-29999	0.3157910	0.29990268	0.1439776	0.1670372
	\$30000-34999	0.3036690	0.28805846	0.1465754	0.1695365
	\$35000-39999	0.3035677	0.28586644	0.1496792	0.1703578
	\$40000-49999	0.2848838	0.26802027	0.1507701	0.1779975
	\$50000-59999	0.2776445	0.24971217	0.1490559	0.1651406
	\$60000-74999	0.2451628	0.20779902	0.1484796	0.1546330
	\$75000-99999	0.2093660	0.16100848	0.1552814	0.1349620
	\$100000-149999	0.1857103	0.13534051	0.1535114	0.1254607
	> \$150000	0.1633755	0.09873015	0.1622744	0.1084418
Housing tenure	Own	0.2089267	0.16031760	0.1550584	0.1801484
	Rent	0.3021760	0.27564949	0.1714583	0.2107407
Housing occupancy	1-2 People	0.2276968	0.18486834	0.1632732	0.1932272
	3-4 People	0.2412469	0.19796330	0.1645403	0.2202286
	5+ People	0.3038374	0.28171474	0.1744384	0.2690233
Building type	Apartment	0.2877216	0.25886065	0.1728960	0.2117330
	Mobile Home	0.2660255	0.24002016	0.1497102	0.1856459
	Single Family Home	0.2221853	0.17259496	0.1623720	0.1982895
Building age	Before 1950	0.2938640	0.25805891	0.1815270	0.2434119

Variable	Category	Mean	Median	SD	IQR
	1950-1959	0.2613181	0.22170204	0.1652544	0.2067353
	1960-1969	0.2688271	0.23169762	0.1674667	0.2168748
	1970-1979	0.2661172	0.22588164	0.1729302	0.2134730
	1980-1989	0.2534525	0.21032089	0.1700812	0.2112252
	1990-1999	0.2093697	0.16542953	0.1495114	0.1883677
	2000-2009	0.1891646	0.14783102	0.1363561	0.1659213
	2010 or later	0.1636207	0.12154126	0.1305500	0.1435179

REFERENCES.

- (1) Gunja, M. Z.; Gumas, E. D.; Williams, R. D. *U.S. Health Care from a Global Perspective, 2022: Accelerating Spending, Worsening Outcomes*. Commonwealth Fund. <https://doi.org/10.26099/8ejy-yc74>.
- (2) Berwick, D. M.; Nolan, T. W.; Whittington, J. The Triple Aim: Care, Health, and Cost. *Health Aff. Proj. Hope* **2008**, 27 (3), 759–769. <https://doi.org/10.1377/hlthaff.27.3.759>.
- (3) Berwick, D. M.; Hackbarth, A. D. Eliminating Waste in US Health Care. *JAMA* **2012**, 307 (14), 1513–1516. <https://doi.org/10.1001/jama.2012.362>.
- (4) Emanuel, E. J.; Fuchs, V. R. The Perfect Storm of Overutilization. *JAMA* **2008**, 299 (23), 2789–2791. <https://doi.org/10.1001/jama.299.23.2789>.
- (5) Bentley, T. G. k.; Effros, R. M.; Palar, K.; Keeler, E. B. Waste in the U.S. Health Care System: A Conceptual Framework. *Milbank Q.* **2008**, 86 (4), 629–659. <https://doi.org/10.1111/j.1468-0009.2008.00537.x>.
- (6) *U.S. Health in International Perspective: Shorter Lives, Poorer Health*; National Academies Press: Washington, D.C., 2013. <https://doi.org/10.17226/13497>.
- (7) National Center for Chronic Disease Prevention and Health Promotion (US) Office on Smoking and Health. *Community Health and Economic Prosperity: Engaging Businesses as Stewards and Stakeholders—A Report of the Surgeon General*; Publications and Reports of the Surgeon General; US Department of Health and Human Services: Washington (DC), 2021.
- (8) Cheung, P. T.; Wiler, J. L.; Lowe, R. A.; Ginde, A. A. National Study of Barriers to Timely Primary Care and Emergency Department Utilization among Medicaid Beneficiaries. *Ann. Emerg. Med.* **2012**, 60 (1), 4-10.e2. <https://doi.org/10.1016/j.annemergmed.2012.01.035>.
- (9) Pappas, G.; Hadden, W. C.; Kozak, L. J.; Fisher, G. F. Potentially Avoidable Hospitalizations: Inequalities in Rates between US Socioeconomic Groups. *Am. J. Public Health* **1997**, 87 (5), 811–816. <https://doi.org/10.2105/AJPH.87.5.811>.
- (10) Lyu, H.; Xu, T.; Brotman, D.; Mayer-Blackwell, B.; Cooper, M.; Daniel, M.; Wick, E. C.; Saini, V.; Brownlee, S.; Makary, M. A. Overtreatment in the United States. *PLoS ONE* **2017**, 12 (9), e0181970. <https://doi.org/10.1371/journal.pone.0181970>.
- (11) *U.S. Health Disadvantage: Causes and Potential Solutions | Policy, Performance, and Evaluation | CDC*. <https://www.cdc.gov/policy/chep/health/index.html> (accessed 2024-09-07).
- (12) Avendano, M.; Glymour, M. M.; Banks, J.; Mackenbach, J. P. Health Disadvantage in US Adults Aged 50 to 74 Years: A Comparison of the Health of Rich and Poor Americans With That of Europeans. *Am. J. Public Health* **2009**, 99 (3), 540–548. <https://doi.org/10.2105/AJPH.2008.139469>.
- (13) Chaityachati, K.; Kangovi, S. Inappropriate ED Visits: Patient Responsibility or an Attribution Bias? *BMJ Qual. Saf.* **2020**, 29 (6), 441–442. <https://doi.org/10.1136/bmjqs-2019-009729>.
- (14) Allen, J. G.; Macomber, J. D. *Healthy Buildings: How Indoor Spaces Can Make You Sick—or Keep You Well*; Harvard University Press, 2022.
- (15) Krieger, N. Discrimination and Health Inequities. *Int. J. Health Serv.* **2014**, 44 (4), 643–710. <https://doi.org/10.2190/HS.44.4.b>.
- (16) Chetty, R.; Stepner, M.; Abraham, S.; Lin, S.; Scuderi, B.; Turner, N.; Bergeron, A.; Cutler, D. The Association Between Income and Life Expectancy in the United States, 2001–2014. *JAMA* **2016**, 315 (16), 1750–1766. <https://doi.org/10.1001/jama.2016.4226>.
- (17) Callaghan, W. M. Overview of Maternal Mortality in the United States. *Semin. Perinatol.* **2012**, 36 (1), 2–6. <https://doi.org/10.1053/j.semperi.2011.09.002>.

- (18) Howell, E. A. Reducing Disparities in Severe Maternal Morbidity and Mortality. *Clin. Obstet. Gynecol.* **2018**, 61 (2), 387–399. <https://doi.org/10.1097/GRF.0000000000000349>.
- (19) MacDorman, M. F.; Thoma, M.; Declercq, E.; Howell, E. A. Racial and Ethnic Disparities in Maternal Mortality in the United States Using Enhanced Vital Records, 2016–2017. *Am. J. Public Health* **2021**, 111 (9), 1673–1681. <https://doi.org/10.2105/AJPH.2021.306375>.
- (20) Trost, S.; Njie, F.; Harvey, A.; Berry, J.; Beauregard, J.; Chandra, G.; Goodman, D. *Pregnancy-Related Deaths: Data From Maternal Mortality Review Committees in 36 U.S. States, 2017–2019*. Centers for Disease Control and Prevention. <https://www.cdc.gov/maternal-mortality/php/data-research/mmrc-2017-2019.html> (accessed 2024-09-06).
- (21) Macias Gil, R.; Marcelin, J. R.; Zuniga-Blanco, B.; Marquez, C.; Mathew, T.; Piggott, D. A. COVID-19 Pandemic: Disparate Health Impact on the Hispanic/Latinx Population in the United States. *J. Infect. Dis.* **2020**, 222 (10), 1592–1595. <https://doi.org/10.1093/infdis/jiaa474>.
- (22) Mackey, K.; Ayers, C. K.; Kondo, K. K.; Saha, S.; Advani, S. M.; Young, S.; Spencer, H.; Rusek, M.; Anderson, J.; Veazie, S.; Smith, M.; Kansagara, D. Racial and Ethnic Disparities in COVID-19–Related Infections, Hospitalizations, and Deaths. *Ann. Intern. Med.* **2021**, 174 (3), 362–373. <https://doi.org/10.7326/M20-6306>.
- (23) Magesh, S.; John, D.; Li, W. T.; Li, Y.; Mattingly-app, A.; Jain, S.; Chang, E. Y.; Ongkeko, W. M. Disparities in COVID-19 Outcomes by Race, Ethnicity, and Socioeconomic Status: A Systematic Review and Meta-Analysis. *JAMA Netw. Open* **2021**, 4 (11), e2134147. <https://doi.org/10.1001/jamanetworkopen.2021.34147>.
- (24) Indian Health Service. *Behavioral Health*. Indian Health Service. <https://www.ihs.gov/newsroom/factsheets/behavioralhealth/> (accessed 2024-09-06).
- (25) Gone, J. P.; Trimble, J. E. American Indian and Alaska Native Mental Health: Diverse Perspectives on Enduring Disparities. *Annu. Rev. Clin. Psychol.* **2012**, 8 (1), 131–160. <https://doi.org/10.1146/annurev-clinpsy-032511-143127>.
- (26) Kwon, S. C.; Kabir, R.; Saadabadi, A. Mental Health Challenges in Caring for American Indians and Alaska Natives. In *StatPearls*; StatPearls Publishing: Treasure Island (FL), 2024.
- (27) Bassett, D.; Buchwald, D.; Manson, S. Posttraumatic Stress Disorder and Symptoms among American Indians and Alaska Natives: A Review of the Literature. *Soc. Psychiatry Psychiatr. Epidemiol.* **2014**, 49 (3), 417–433. <https://doi.org/10.1007/s00127-013-0759-y>.
- (28) Stone, D.; Trinh, E.; Zhou, H.; Welder, L.; End of Horn, P.; Fowler, K.; Ivey-Stephenson, A. Suicides Among American Indian or Alaska Native Persons — National Violent Death Reporting System, United States, 2015–2020. *Cent. Dis. Control Prev.* **2022**, 71. <https://doi.org/10.15585/mmwr.mm7137a1>.
- (29) Schweizer, E. *Environmental Justice: An Interview with Robert Bullard*. <https://ejnet.org/ej/bullard.html> (accessed 2024-09-06).
- (30) Bailey, Z. D.; Krieger, N.; Agénor, M.; Graves, J.; Linos, N.; Bassett, M. T. Structural Racism and Health Inequities in the USA: Evidence and Interventions. *Lancet Lond. Engl.* **2017**, 389 (10077), 1453–1463. [https://doi.org/10.1016/S0140-6736\(17\)30569-X](https://doi.org/10.1016/S0140-6736(17)30569-X).
- (31) Bullard, R. D. Robert Bullard: Dumping in Dixie: Race, Class and Environmental Quality. *Ecol. Law Q.* **1992**, 19 (3), 591–609.
- (32) Sze, J. Environmental Justice in a Moment of Danger. *ISLE Interdiscip. Stud. Lit. Environ.* **2020**, 27 (4), 890–891. <https://doi.org/10.1093/isle/isaa166>.
- (33) Rothstein, R. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Economic Policy Institute. <https://www.epi.org/publication/the-color-of-law-a-forgotten-history-of-how-our-government-segregated-america/> (accessed 2024-09-06).

- (34) Massey, D. S.; Rothwell, J.; Domina, T. The Changing Bases of Segregation in the United States. *Ann. Am. Acad. Pol. Soc. Sci.* **2009**, 626 (1), 10.1177/0002716209343558. <https://doi.org/10.1177/0002716209343558>.
- (35) Mohai, P.; Pellow, D.; Roberts, J. T. *Environmental Justice | Annual Reviews*. <https://www.annualreviews.org/content/journals/10.1146/annurev-environ-082508-094348> (accessed 2024-09-06).
- (36) Bullard, R. D. *The Quest for Environmental Justice*; 2015.
- (37) White, K.; Borrell, L. N. Racial/Ethnic Residential Segregation: Framing the Context of Health Risk and Health Disparities. *Health Place* **2011**, 17 (2), 438–448. <https://doi.org/10.1016/j.healthplace.2010.12.002>.
- (38) Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care*; Smedley, B. D., Stith, A. Y., Nelson, A. R., Eds.; National Academies Press (US): Washington (DC), 2003.
- (39) Collins, C. A.; Williams, D. R. Segregation and Mortality: The Deadly Effects of Racism? *Sociol. Forum* **1999**, 14 (3), 495–523. <https://doi.org/10.1023/A:1021403820451>.
- (40) Krieger, N.; Chen, J. T.; Coull, B. A.; Beckfield, J.; Kiang, M. V.; Waterman, P. D. Jim Crow and Premature Mortality Among the US Black and White Population, 1960–2009. *Epidemiol. Camb. Mass* **2014**, 25 (4), 494–504. <https://doi.org/10.1097/EDE.0000000000000104>.
- (41) Williams, D. R.; Collins, C. *Racial Residential Segregation: A Fundamental Cause of Racial Disparities in Health*. <https://doi.org/10.1093/phr/116.5.404>.
- (42) Acevedo-Garcia, D.; Lochner, K. A.; Osypuk, T. L.; Subramanian, S. V. Future Directions in Residential Segregation and Health Research: A Multilevel Approach. *Am. J. Public Health* **2003**, 93 (2), 215–221. <https://doi.org/10.2105/AJPH.93.2.215>.
- (43) Kramer, M. R.; Hogue, C. R. Is Segregation Bad for Your Health? *Epidemiol. Rev.* **2009**, 31 (1), 178–194. <https://doi.org/10.1093/epirev/mxp001>.
- (44) Landrine, H.; Corral, I. Separate and Unequal: Residential Segregation and Black Health Disparities. *Ethn. Dis.* **2009**, 19 (2), 179–184.
- (45) Swope, C. B.; Hernández, D.; Cushing, L. J. The Relationship of Historical Redlining with Present-Day Neighborhood Environmental and Health Outcomes: A Scoping Review and Conceptual Model. *J. Urban Health* **2022**, 99 (6), 959–983. <https://doi.org/10.1007/s11524-022-00665-z>.
- (46) Rao, N. D.; Sauer, P.; Gidden, M.; Riahi, K. Income Inequality Projections for the Shared Socioeconomic Pathways (SSPs). *Futures* **2019**, 105, 27–39. <https://doi.org/10.1016/j.futures.2018.07.001>.
- (47) U.S. Department of Commerce. *Geographic Inequality on the Rise in the U.S.* U.S. Department of Commerce. <https://www.commerce.gov/news/blog/2023/06/geographic-inequality-rise-us> (accessed 2024-09-12).
- (48) Madhav, N.; Oppenheim, B.; Gallivan, M.; Mulembakani, P.; Rubin, E.; Wolfe, N. Pandemics: Risks, Impacts, and Mitigation. In *Disease Control Priorities: Improving Health and Reducing Poverty*; Jamison, D. T., Gelband, H., Horton, S., Jha, P., Laxminarayan, R., Mock, C. N., Nugent, R., Eds.; The International Bank for Reconstruction and Development / The World Bank: Washington (DC), 2017.
- (49) KC, B.; Shepherd, J. M.; King, A. W.; Johnson Gaither, C. Multi-Hazard Climate Risk Projections for the United States. *Nat. Hazards* **2021**, 105 (2), 1963–1976. <https://doi.org/10.1007/s11069-020-04385-y>.

- (50) Dunlap, R. E.; Brulle, R. J. *Climate Change and Society: Sociological Perspectives*; Oxford University Press, 2015.
- (51) Naylor, A.; Ford, J.; Pearce, T.; Van Alstine, J. Conceptualizing Climate Vulnerability in Complex Adaptive Systems. *One Earth* **2020**, 2 (5), 444–454. <https://doi.org/10.1016/j.oneear.2020.04.011>.
- (52) *Housing Markets and Climate Migration* | Urban Institute. <https://www.urban.org/research/publication/housing-markets-and-climate-migration> (accessed 2024-09-12).
- (53) Tanton, R. Spatial Microsimulation: Developments and Potential Future Directions. *Int. J. Microsimulation* **2017**, 11 (1), 143–161. <https://doi.org/10.34196/ijm.00176>.
- (54) Lovelace, R.; Dumont, M. *Spatial Microsimulation with R*.
- (55) *Geography Matters: Simulating the Local Impacts of National Social Policies*; Ballas, D., Joseph Rowntree Foundation, Eds.; Contemporary research issues; Joseph Rowntree Foundation: York, 2005.
- (56) Krieger, N. Theories for Social Epidemiology in the 21st Century: An Ecosocial Perspective. *Int. J. Epidemiol.* **2001**, 30 (4), 668–677. <https://doi.org/10.1093/ije/30.4.668>.
- (57) Krieger, N. *Epidemiology and the People's Health: Theory and Context, Second Edition*, 2nd edn.; Oxford Academic, 2024.
- (58) Spatial Epidemiology: Methods and Applications. In *Spatial Epidemiology: Methods and Applications*; Elliott, P., Wakefield, J. C., Best, N. G., Briggs, D. J., Elliott, P., Wakefield, J., Best, N., Briggs, D., Eds.; Oxford University Press, 2001; p 0. <https://doi.org/10.1093/acprof:oso/9780198515326.003.0001>.
- (59) Elliott, P.; Wartenberg, D. Spatial Epidemiology: Current Approaches and Future Challenges. *Environ. Health Perspect.* **2004**, 112 (9), 998–1006. <https://doi.org/10.1289/ehp.6735>.
- (60) Kirby, R. S.; Delmelle, E.; Eberth, J. M. Advances in Spatial Epidemiology and Geographic Information Systems. *Ann. Epidemiol.* **2017**, 27 (1), 1–9. <https://doi.org/10.1016/j.annepidem.2016.12.001>.
- (61) Morrison, C. N.; Mair, C. F.; Bates, L.; Duncan, D. T.; Branas, C. C.; Bushover, B. R.; Mehranbod, C. A.; Gobaud, A. N.; Uong, S.; Forrest, S.; Roberts, L.; Rundle, A. G. Defining Spatial Epidemiology: A Systematic Review and Re-Orientation. *Epidemiol. Camb. Mass* **2024**, 35 (4), 542–555. <https://doi.org/10.1097/EDE.0000000000001738>.
- (62) Lovelace, R. The Energy Costs of Commuting: A Spatial Microsimulation Approach, 2014.
- (63) Openshaw, S. *The Modifiable Areal Unit Problem*; Concepts and techniques in modern geography; Geo Books: Norwich [Norfolk], 1983.
- (64) Macintyre, S.; Ellaway, A.; Cummins, S. Place Effects on Health: How Can We Conceptualise, Operationalise and Measure Them? *Soc. Sci. Med.* **2002**, 55 (1), 125–139. [https://doi.org/10.1016/S0277-9536\(01\)00214-3](https://doi.org/10.1016/S0277-9536(01)00214-3).
- (65) Macintyre, S.; Ellaway, A. Ecological Approaches: Rediscovering the Role of the Physical and Social Environment. In *Social Epidemiology*; Berkman, L. F., Ph. D., Kawachi, I., M. D., Ph. D., Eds.; Oxford University Press, 2000; p 0. <https://doi.org/10.1093/oso/9780195083316.003.0014>.
- (66) Pearce, N. *The ecological fallacy strikes back* | *Journal of Epidemiology & Community Health*. <https://jech.bmj.com/content/54/5/326.short> (accessed 2024-09-07).
- (67) Schwartz, S. The Fallacy of the Ecological Fallacy: The Potential Misuse of a Concept and the Consequences. *Am. J. Public Health* **1994**, 84 (5), 819–824. <https://doi.org/10.2105/AJPH.84.5.819>.

- (68) Robinson, W. Ecological Correlations and the Behavior of Individuals*. *Int. J. Epidemiol.* **2009**, *38* (2), 337–341. <https://doi.org/10.1093/ije/dyn357>.
- (69) Greenland, S.; Robins, J. Invited Commentary: Ecologic Studies—Biases, Misconceptions, and Counterexamples. *Am. J. Epidemiol.* **1994**, *139* (8), 747–760. <https://doi.org/10.1093/oxfordjournals.aje.a117069>.
- (70) CDC/ATSDR Social Vulnerability Index (CDC/ATSDR SVI). <https://www.atsdr.cdc.gov/placeandhealth/svi/index.html> (accessed 2024-09-07).
- (71) CDC. Environmental Justice Index (EJI). Centers for Disease Control and Prevention. <https://www.atsdr.cdc.gov/placeandhealth/eji/index.html> (accessed 2024-09-07).
- (72) Climate and Economic Justice Screening Tool. Climate and Economic Justice Screening Tool. <https://screeningtool.geoplatform.gov> (accessed 2024-09-07).
- (73) US EPA, O. EJScreen: Environmental Justice Screening and Mapping Tool. <https://www.epa.gov/ejscreen> (accessed 2024-09-07).
- (74) Resilience Analysis and Planning Tool (RAPT) | FEMA.gov. <https://www.fema.gov/about/reports-and-data/resilience-analysis-planning-tool> (accessed 2024-09-07).
- (75) Diderichsen, F.; Hallqvist, J.; Whitehead, M. Differential Vulnerability and Susceptibility: How to Make Use of Recent Development in Our Understanding of Mediation and Interaction to Tackle Health Inequalities. *Int. J. Epidemiol.* **2019**, *48* (1), 268–274. <https://doi.org/10.1093/ije/dyy167>.
- (76) Krieger, N. Health Equity and the Fallacy of Treating Causes of Population Health as If They Sum to 100%. *Am. J. Public Health* **2017**, *107* (4), 541–549. <https://doi.org/10.2105/AJPH.2017.303655>.
- (77) Ballas, D.; Clarke, G.; Hynes, S.; Lennon, J.; Morrissey, K.; O'Donoghue, C. A Review of Microsimulation for Policy Analysis. In *Spatial Microsimulation for Rural Policy Analysis*; O'Donoghue, C., Ballas, D., Clarke, G., Hynes, S., Morrissey, K., Eds.; Springer: Berlin, Heidelberg, 2013; pp 35–54. https://doi.org/10.1007/978-3-642-30026-4_3.
- (78) O'Donoghue, C.; Morrissey, K.; Lennon, J. *Spatial microsimulation modelling: A review of applications and methodological choices* | *International Journal of Microsimulation*. <https://microsimulation.pub/articles/00093> (accessed 2024-09-07).
- (79) Mitchell, R.; Dorling, D.; Shaw, M. Population Production and Modelling Mortality—an Application of Geographic Information Systems in Health Inequalities Research. *Health Place* **2002**, *8* (1), 15–24. [https://doi.org/10.1016/S1353-8292\(01\)00032-6](https://doi.org/10.1016/S1353-8292(01)00032-6).
- (80) R: The R Project for Statistical Computing. <https://www.r-project.org/> (accessed 2024-09-06).
- (81) FAS Research Computing. FAS Research Computing. <https://www.rc.fas.harvard.edu/> (accessed 2024-09-06).
- (82) Lovelace, R.; Ballas, D.; Watson, M. A Spatial Microsimulation Approach for the Analysis of Commuter Patterns: From Individual to Regional Levels. *J. Transp. Geogr.* **2014**, *34*, 282–296. <https://doi.org/10.1016/j.jtrangeo.2013.07.008>.
- (83) Tanton, R. A Review of Spatial Microsimulation Methods. *Int. J. Microsimulation* **2014**, *7* (1), 4–25.
- (84) Deming, W. E.; Stephan, F. F. On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals Are Known. *Ann. Math. Stat.* **1940**, *11* (4), 427–444.
- (85) Mosteller, F. Association and Estimation in Contingency Tables. *J. Am. Stat. Assoc.* **1968**, *63* (321), 1–28. <https://doi.org/10.2307/2283825>.
- (86) Fienberg, S. E. An Iterative Procedure for Estimation in Contingency Tables. *Ann. Math. Stat.* **1970**, *41* (3), 907–917. <https://doi.org/10.1214/aoms/1177696968>.

- (87) Wong, D. W. S. The Reliability of Using the Iterative Proportional Fitting Procedure. *Prof. Geogr.* **1992**, *44* (3), 340–348. <https://doi.org/10.1111/j.0033-0124.1992.00340.x>.
- (88) Birkin, M.; Clarke, M. Synthesis—A Synthetic Spatial Information System for Urban and Regional Analysis: Methods and Examples. *Environ. Plan. Econ. Space* **1988**, *20* (12), 1645–1671. <https://doi.org/10.1068/a201645>.
- (89) Birkin, M.; Clarke, M. The Generation of Individual and Household Incomes at the Small Area Level Using Synthesis. *Reg. Stud.* **1989**. <https://doi.org/10.1080/00343408912331345702>.
- (90) Williamson, P.; Birkin, M.; Rees, P. H. The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records. *Environ. Plan. A* **1998**, *30* (5), 785–816. <https://doi.org/10.1068/a300785>.
- (91) Lovelace, R.; Ballas, D. ‘Truncate, Replicate, Sample’: A Method for Creating Integer Weights for Spatial Microsimulation. *Comput. Environ. Urban Syst.* **2013**, *41*, 1–11. <https://doi.org/10.1016/j.compenvurbsys.2013.03.004>.
- (92) Smith, A.; Lovelace, R.; Birkin, M. Population Synthesis with Quasirandom Integer Sampling. *J. Artif. Soc. Soc. Simul.* **2017**, *20* (4), 14.
- (93) Norman, P. Putting Iterative Proportional Fitting on the Researcher’s Desk. *Prof. Geogr.* **2016**, *68* (3), 451–461. <https://doi.org/10.1080/00330124.2015.1099449>.
- (94) Pritchard, D. R.; Miller, E. J. Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously. *Transportation* **2012**, *39* (3), 685–704. <https://doi.org/10.1007/s11116-011-9367-4>.
- (95) Tanton, R.; Edwards, K. Introduction to Spatial Microsimulation: History, Methods and Applications. In *Spatial microsimulation: a reference guide for users*; Tanton, R., Edwards, K. L., Eds.; Springer: Netherlands, 2013; Vol. 6, pp 3–8. https://doi.org/10.1007/978-94-007-4623-7_1.
- (96) *Spatial Microsimulation: A Reference Guide for Users*; Tanton, R., Edwards, K., Eds.; Springer Netherlands: Dordrecht, 2013. <https://doi.org/10.1007/978-94-007-4623-7>.
- (97) Burden, S.; Steel, D. Constraint Choice for Spatial Microsimulation. *Popul. Space Place* **2016**, *22* (6), 568–583. <https://doi.org/10.1002/psp.1942>.
- (98) Jacobs, D. E.; Wilson, J.; Dixon, S. L.; Smith, J.; Evens, A. The Relationship of Housing and Population Health: A 30-Year Retrospective Analysis. *Environ. Health Perspect.* **2009**, *117* (4), 597–604. <https://doi.org/10.1289/ehp.0800086>.
- (99) Adamkiewicz, G.; Zota, A. R.; Fabian, M. P.; Chahine, T.; Julien, R.; Spengler, J. D.; Levy, J. I. Moving Environmental Justice Indoors: Understanding Structural Influences on Residential Exposure Patterns in Low-Income Communities. *Am. J. Public Health* **2011**, *101* (Suppl 1), S238–S245. <https://doi.org/10.2105/AJPH.2011.300119>.
- (100) Kawachi, I.; Kennedy, B. P. Socioeconomic Determinants of Health : Health and Social Cohesion: Why Care about Income Inequality? *BMJ* **1997**, *314* (7086), 1037. <https://doi.org/10.1136/bmj.314.7086.1037>.
- (101) Massey, D. S. American Apartheid: Segregation and the Making of the Underclass. *Am. J. Sociol.* **1990**, *96* (2), 329–357.
- (102) Massey, D. S.; Denton, N. A. The Dimensions of Residential Segregation*. *Soc. Forces* **1988**, *67* (2), 281–315. <https://doi.org/10.1093/sf/67.2.281>.
- (103) Diez Roux, A. V. Investigating Neighborhood and Area Effects on Health. *Am. J. Public Health* **2001**, *91* (11), 1783–1789. <https://doi.org/10.2105/AJPH.91.11.1783>.
- (104) Diez Roux, A. V.; Mair, C. Neighborhoods and Health. *Ann. N. Y. Acad. Sci.* **2010**, *1186* (1), 125–145. <https://doi.org/10.1111/j.1749-6632.2009.05333.x>.
- (105) *Neighborhood and Life Chances: How Place Matters in Modern America*; Newburger, H. B., Birch, E. L., Wachter, S. M., Series Eds.; University of Pennsylvania Press, 2011.

- (106) Halfon, N.; Hochstein, M. Life Course Health Development: An Integrated Framework for Developing Health, Policy, and Research. *Milbank Q.* **2002**, *80* (3), 433–479. <https://doi.org/10.1111/1468-0009.00019>.
- (107) Mauvais-Jarvis, F.; Merz, N. B.; Barnes, P. J.; Brinton, R. D.; Carrero, J.-J.; DeMeo, D. L.; Vries, G. J. D.; Epperson, C. N.; Govindan, R.; Klein, S. L.; Lonardo, A.; Maki, P. M.; McCullough, L. D.; Regitz-Zagrosek, V.; Regensteiner, J. G.; Rubin, J. B.; Sandberg, K.; Suzuki, A. Sex and Gender: Modifiers of Health, Disease, and Medicine. *The Lancet* **2020**, *396* (10250), 565–582. [https://doi.org/10.1016/S0140-6736\(20\)31561-0](https://doi.org/10.1016/S0140-6736(20)31561-0).
- (108) Schwartz, J.; Bellinger, D.; Glass, T. Expanding the Scope of Environmental Risk Assessment to Better Include Differential Vulnerability and Susceptibility. *Am. J. Public Health* **2011**, *101* (Suppl 1), S88–S93. <https://doi.org/10.2105/AJPH.2011.300268>.
- (109) Bureau, U. C. *Glossary*. Census.gov. <https://www.census.gov/programs-surveys/geography/about/glossary.html> (accessed 2024-09-06).
- (110) Jarmin, R. S. Disclosure Avoidance for the 2020 Census: An Introduction.
- (111) Abowd, J. M.; Ashmead, R.; Cumings-Menon, R.; Garfinkel, S.; Heineck, M.; Heiss, C.; Johns, R.; Kifer, D.; Leclerc, P.; Machanavajjhala, A.; Moran, B.; Sexton, W.; Spence, M.; Zhuravlev, P. The 2020 Census Disclosure Avoidance System TopDown Algorithm. *Harv. Data Sci. Rev.* **2022**, No. Special Issue 2. <https://doi.org/10.1162/99608f92.529e3cb9>.
- (112) Lewis, P. G. T.; Chiu, W. A.; Nasser, E.; Proville, J.; Barone, A.; Danforth, C.; Kim, B.; Prozzi, J.; Craft, E. Characterizing Vulnerabilities to Climate Change Across the United States. *Environ. Int.* **2023**, *172*, 107772. <https://doi.org/10.1016/j.envint.2023.107772>.
- (113) Krieger, N.; Chen, J. T.; Waterman, P. D.; Soobader, M.-J.; Subramanian, S. V.; Carson, R. Geocoding and Monitoring of US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does the Choice of Area-Based Measure and Geographic Level Matter?: The Public Health Disparities Geocoding Project. *Am. J. Epidemiol.* **2002**, *156* (5), 471–482. <https://doi.org/10.1093/aje/kwf068>.
- (114) US EPA, O. *Air Quality System (AQS)*. <https://www.epa.gov/aqs> (accessed 2024-09-07).
- (115) US EPA, O. *Toxics Release Inventory (TRI) Program*. <https://www.epa.gov/toxics-release-inventory-tri-program> (accessed 2024-09-07).
- (116) US EPA, O. *EnviroAtlas*. <https://www.epa.gov/enviroatlas> (accessed 2024-09-07).
- (117) *National Risk Index for Natural Hazards | FEMA.gov*. <https://www.fema.gov/flood-maps/products-tools/national-risk-index> (accessed 2024-09-07).
- (118) US Department of Commerce, N. *Climate*. <https://www.weather.gov/wrh/climate> (accessed 2024-09-07).
- (119) Bureau, U. C. *Relationship Files*. Census.gov. <https://www.census.gov/geographies/reference-files/time-series/geo/relationship-files.html> (accessed 2024-09-06).
- (120) Bureau, U. C. *Code Lists, Definitions, and Accuracy*. Census.gov. <https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html> (accessed 2024-09-06).
- (121) *TigerWEB*. esri. <https://tigerweb.geo.census.gov/tigerweb/> (accessed 2024-09-06).
- (122) Bureau, U. C. *2022 ACS PUMS Estimates*. Census.gov. <https://www.census.gov/programs-surveys/acs/microdata/documentation/2022.html> (accessed 2024-09-06).
- (123) Lovelace, R.; Birkin, M.; Ballas, D.; Van Leeuwen, E. Evaluating the Performance of Iterative Proportional Fitting for Spatial Microsimulation: New Tests for an Established Technique. *J. Artif. Soc. Soc. Simul.* **2015**, *18* (2), 21. <https://doi.org/10.18564/jasss.2768>.
- (124) Hajat, A.; MacLehose, R. F.; Rosofsky, A.; Walker, K. D.; Clougherty, J. E. Confounding by Socioeconomic Status in Epidemiological Studies of Air Pollution and Health: Challenges and

- Opportunities. *Environ. Health Perspect.* **2021**, *129* (6), 065001. <https://doi.org/10.1289/EHP7980>.
- (125) Tomintz, M. N.; Clarke, G. P.; Rigby, J. E. The Geography of Smoking in Leeds: Estimating Individual Smoking Rates and the Implications for the Location of Stop Smoking Services. *Area* **2008**, *40* (3), 341–353. <https://doi.org/10.1111/j.1475-4762.2008.00837.x>.
 - (126) Tomintz, M.; Kosar, B.; Clarke, G. smokeSALUD: Exploring the Effect of Demographic Change on the Smoking Prevalence at Municipality Level in Austria. *Int. J. Health Geogr.* **2016**, *15* (1), 36. <https://doi.org/10.1186/s12942-016-0066-4>.
 - (127) Smith, D. M.; Pearce, J. R.; Harland, K. Can a Deterministic Spatial Microsimulation Model Provide Reliable Small-Area Estimates of Health Behaviours? An Example of Smoking Prevalence in New Zealand. *Health Place* **2011**, *17* (2), 618–624. <https://doi.org/10.1016/j.healthplace.2011.01.001>.
 - (128) Tomintz, M. N.; Barnett, R. Geosimulation Approach for Filling the Gap of Non-Response Smoking Data from the Census 2013: A Spatial Analysis of Census Area Unit Geographies. *N. Z. Geogr.* **2018**, *74* (3), 152–159. <https://doi.org/10.1111/nzg.12199>.
 - (129) James, W. H. M.; Lomax, N.; Birkin, M. Local Level Estimates of Food, Drink and Tobacco Expenditure for Great Britain. *Sci. Data* **2019**, *6* (1), 56. <https://doi.org/10.1038/s41597-019-0064-z>.
 - (130) Hermes, K.; Poulsen, M. Small Area Estimates of Smoking Prevalence in London. Testing the Effect of Input Data. *Health Place* **2012**, *18* (3), 630–638. <https://doi.org/10.1016/j.healthplace.2011.12.010>.
 - (131) Smith, D. M.; Harland, K.; Clarke, G. P. SimHealth: Estimating Small Area Populations Using Deterministic Spatial Microsimulation in Leeds and Bradford.
 - (132) Crespo, R.; Alvarez, C.; Hernandez, I.; García, C. A Spatially Explicit Analysis of Chronic Diseases in Small Areas: A Case Study of Diabetes in Santiago, Chile. *Int. J. Health Geogr.* **2020**, *19*, 24. <https://doi.org/10.1186/s12942-020-00217-1>.
 - (133) Edwards, K. L.; Clarke, G. P. The Design and Validation of a Spatial Microsimulation Model of Obesogenic Environments for Children in Leeds, UK: SimObesity. *Soc. Sci. Med.* **2009**, *69* (7), 1127–1134. <https://doi.org/10.1016/j.socscimed.2009.07.037>.
 - (134) Koh, K.; Grady, S. C.; Darden, J. T.; Vojnovic, I. Adult Obesity Prevalence at the County Level in the United States, 2000–2010: Downscaling Public Health Survey Data Using a Spatial Microsimulation Approach. *Spat. Spatio-Temporal Epidemiol.* **2018**, *26*, 153–164. <https://doi.org/10.1016/j.sste.2017.10.001>.
 - (135) Cataife, G. Small Area Estimation of Obesity Prevalence and Dietary Patterns: A Model Applied to Rio de Janeiro City, Brazil. *Health Place* **2014**, *26*, 47–52. <https://doi.org/10.1016/j.healthplace.2013.12.004>.
 - (136) Spooner, F.; Abrams, J. F.; Morrissey, K.; Shaddick, G.; Batty, M.; Milton, R.; Dennett, A.; Lomax, N.; Malleson, N.; Nelissen, N.; Coleman, A.; Nur, J.; Jin, Y.; Greig, R.; Shenton, C.; Birkin, M. A Dynamic Microsimulation Model for Epidemics. *Soc. Sci. Med.* **2021**, *291*, 114461. <https://doi.org/10.1016/j.socscimed.2021.114461>.
 - (137) Xu, Z.; Glass, K.; Lau, C. L.; Geard, N.; Graves, P.; Clements, A. A Synthetic Population for Modelling the Dynamics of Infectious Disease Transmission in American Samoa. *Sci. Rep.* **2017**, *7*, 16725. <https://doi.org/10.1038/s41598-017-17093-8>.
 - (138) Morrissey, K.; Hynes, S.; Clarke, G.; O'Donoghue, C. Examining the Factors Associated with Depression at the Small Area Level in Ireland Using Spatial Microsimulation Techniques. *Ir. Geogr.* **2010**, *43* (1), 1–22. <https://doi.org/10.1080/00750771003696489>.

- (139) Wiki, J.; Marek, L.; Sibley, C.; Exeter, D. Estimating Quality of Life: A Spatial Microsimulation Model of Well-Being in Aotearoa New Zealand. *Soc. Sci. Med.* **2023**, *330*, 116054. <https://doi.org/10.1016/j.socscimed.2023.116054>.
- (140) Riva, M.; Smith, D. M. Generating Small-Area Prevalence of Psychological Distress and Alcohol Consumption: Validation of a Spatial Microsimulation Method. *Soc. Psychiatry Psychiatr. Epidemiol.* **2012**, *47* (5), 745–755. <https://doi.org/10.1007/s00127-011-0376-6>.
- (141) Markham, F.; Young, M.; Doran, B. Improving Spatial Microsimulation Estimates of Health Outcomes by Including Geographic Indicators of Health Behaviour: The Example of Problem Gambling. *Health Place* **2017**, *46*, 29–36. <https://doi.org/10.1016/j.healthplace.2017.04.008>.
- (142) Morrissey, K.; Ballas, D.; Clarke, G.; Hynes, S.; O'Donoghue, C. Spatial Access to Health Services. *Adv. Spat. Sci.* **2013**, 213–230.
- (143) Morrissey, K.; Clarke, G.; Ballas, D.; Hynes, S.; O'Donoghue, C. Examining Access to GP Services in Rural Ireland Using Microsimulation Analysis. *Area* **2008**, *40* (3), 354–364. <https://doi.org/10.1111/j.1475-4762.2008.00844.x>.
- (144) Tomintz, M. N.; Clarke, G. P.; Rigby, J. E.; Green, J. M. Optimising the Location of Antenatal Classes. *Midwifery* **2013**, *29* (1), 33–43. <https://doi.org/10.1016/j.midw.2011.10.010>.
- (145) Smith, D. M.; Clarke, G. P.; Ransley, J.; Cade, J. Food Access and Health : A Microsimulation Framework for Analysis. *Stud. Reg. Sci.* **2006**, *35* (4), 909–927. <https://doi.org/10.2457/srs.35.909>.
- (146) Symonds, P.; Hutchinson, E.; Ibbetson, A.; Taylor, J.; Milner, J.; Chalabi, Z.; Davies, M.; Wilkinson, P. MicroEnv: A Microsimulation Model for Quantifying the Impacts of Environmental Policies on Population Health and Health Inequalities. *Sci. Total Environ.* **2019**, *697*, 134105. <https://doi.org/10.1016/j.scitotenv.2019.134105>.
- (147) Levy, J. I.; Fabian, M. P.; Peters, J. L. Community-Wide Health Risk Assessment Using Geographically Resolved Demographic Data: A Synthetic Population Approach. *PLoS ONE* **2014**, *9* (1), e87144. <https://doi.org/10.1371/journal.pone.0087144>.
- (148) Tomintz, M. N.; Kosar, B.; García-Barrios, V. M. simSALUD: Design and Implementation of an Open-Source Wizard based Spatial Microsimulation Framework. *Int. J. Microsimulation* **2016**, *10* (2), 118–143. <https://doi.org/10.34196/ijm.00159>.
- (149) *Low-Income Energy Affordability Data (LEAD) Tool and Community Energy Solutions*. Energy.gov. <https://www.energy.gov/scep/low-income-energy-affordability-data-lead-tool-and-community-energy-solutions> (accessed 2024-09-07).
- (150) Kraev, T. A.; Adamkiewicz, G.; Hammond, S. K.; Spengler, J. D. Indoor Concentrations of Nicotine in Low-Income, Multi-Unit Housing: Associations with Smoking Behaviours and Housing Characteristics. *Tob. Control* **2009**, *18* (6), 438–444. <https://doi.org/10.1136/tc.2009.029728>.
- (151) King, B. A.; Travers, M. J.; Cummings, K. M.; Mahoney, M. C.; Hyland, A. J. Secondhand Smoke Transfer in Multiunit Housing. *Nicotine Tob. Res. Off. J. Soc. Res. Nicotine Tob.* **2010**, *12* (11), 1133–1141. <https://doi.org/10.1093/ntr/ntq162>.
- (152) Alhasan, D. M.; Gaston, S. A.; Gullett, L. R.; Braxton Jackson, W.; Stanford, F. C.; Jackson, C. L. Neighborhood Social Cohesion and Obesity in the United States. *Endocr. Metab. Sci.* **2023**, *11*, 100129. <https://doi.org/10.1016/j.endmts.2023.100129>.
- (153) Howell, N. A.; Booth, G. L. The Weight of Place: Built Environment Correlates of Obesity and Diabetes. *Endocr. Rev.* **2022**, *43* (6), 966–983. <https://doi.org/10.1210/endrev/bnac005>.
- (154) Papas, M. A.; Alberg, A. J.; Ewing, R.; Helzlsouer, K. J.; Gary, T. L.; Klassen, A. C. The Built Environment and Obesity. *Epidemiol. Rev.* **2007**, *29* (1), 129–143. <https://doi.org/10.1093/epirev/mxm009>.

- (155) Baik, S.; Hines, J. F.; Sim, J. Racial Disparities in the Energy Burden beyond Socio-Economic Inequality. *Energy Econ.* **2023**, *127*, 107098. <https://doi.org/10.1016/j.eneco.2023.107098>.
- (156) Ballas, D.; Clarke, G.; Dorling, D.; Eyre, H.; Thomas, B.; Rossiter, D. SimBritain: A Spatial Microsimulation Approach to Population Dynamics. *Popul. Space Place* **2005**, *11* (1), 13–34. <https://doi.org/10.1002/psp.351>.
- (157) Campbell, M.; Ballas, D. SimAlba: A Spatial Microsimulation Approach to the Analysis of Health Inequalities. *Front. Public Health* **2016**, *4*. <https://doi.org/10.3389/fpubh.2016.00230>.
- (158) Wheaton, W. D.; Cajka, J. C.; Chasteen, B. M.; Wagener, D. K.; Cooley, P. C.; Ganapathi, L.; Roberts, D. J.; Allpress, J. L. Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. *Methods Rep. RTI Press* **2009**, *2009* (10), 905. <https://doi.org/10.3768/rtpress.2009.mr.0010.0905>.
- (159) Beckman, R. J.; Baggerly, K. A.; McKay, M. D. Creating Synthetic Baseline Populations. *Transp. Res. Part Policy Pract.* **1996**, *30* (6), 415–429. [https://doi.org/10.1016/0965-8564\(96\)00004-3](https://doi.org/10.1016/0965-8564(96)00004-3).
- (160) Cajka, J. C.; Cooley, P. C.; Wheaton, W. D. Attribute Assignment to a Synthetic Population in Support of Agent-Based Disease Modeling. *Methods Rep. RTI Press* **2010**, *19* (1009), 1–14. <https://doi.org/10.3768/rtpress.2010.mr.0019.1009>.
- (161) Grefenstette, J. J.; Brown, S. T.; Rosenfeld, R.; DePasse, J.; Stone, N. T.; Cooley, P. C.; Wheaton, W. D.; Fyshe, A.; Galloway, D. D.; Sriram, A.; Guclu, H.; Abraham, T.; Burke, D. S. FRED (A Framework for Reconstructing Epidemic Dynamics): An Open-Source Software System for Modeling Infectious Diseases and Control Strategies Using Census-Based Populations. *BMC Public Health* **2013**, *13*, 940. <https://doi.org/10.1186/1471-2458-13-940>.
- (162) *COVID-19 Data Insights Tool* | RTI. <https://www.rti.org/insights/covid-19-data-insights-tool> (accessed 2024-09-07).
- (163) *Infectious Disease Modeling: NC MInD Program* | RTI. <https://www.rti.org/insights/north-carolina-modeling-infectious-diseases-nc-mind-program-our-innovative-modeling-plan> (accessed 2024-09-07).
- (164) Hadley, E.; Rhea, S.; Jones, K.; Li, L.; Stoner, M.; Bobashev, G. Enhancing the Prediction of Hospitalization from a COVID-19 Agent-Based Model: A Bayesian Method for Model Parameter Estimation. *PLoS ONE* **2022**, *17* (3), e0264704. <https://doi.org/10.1371/journal.pone.0264704>.
- (165) Jones, K.; Hadley, E.; Kery, C.; Preiss, A.; Stoner, M. C. D.; Rhea, S. Agent-Based Model Framework for the North Carolina Modeling Infectious Diseases Program (NC MInD ABM) Overview, Design Concepts, and Details Protocol. arXiv February 25, 2022. <https://doi.org/10.48550/arXiv.2202.06853>.
- (166) Jones, K.; Hadley, E.; Preiss, S.; Lofgren, E. T.; Rice, D. P.; Stoner, M. C. D.; Rhea, S.; Adams, J. W. Estimate of Undetected Severe Acute Respiratory Coronavirus Virus 2 (SARS-CoV-2) Infection in Acute-Care Hospital Settings Using an Individual-Based Microsimulation Model. *Infect. Control Hosp. Epidemiol.* **2023**, *44* (6), 898–907. <https://doi.org/10.1017/ice.2022.174>.
- (167) Adams, J. W.; Jones, K.; Preiss, S.; Hadley, E.; Segelman, M. Evaluating Policies to Decrease the Risk of Introducing SARS-CoV-2 Infections to Nursing Home Facilities. *J. Appl. Gerontol.* **2023**, *42* (7), 1505–1516. <https://doi.org/10.1177/07334648231155873>.
- (168) Lee, B. Y.; Brown, S. T.; Cooley, P.; Potter, M. A.; Wheaton, W. D.; Voorhees, R. E.; Stebbins, S.; Grefenstette, J. J.; Zimmer, S. M.; Zimmerman, R.; Assi, T.-M.; Bailey, R. R.; Wagener, D. K.; Burke, D. S. Simulating School Closure Strategies to Mitigate an Influenza Epidemic. *J. Public Health Manag. Pract. JPHMP* **2010**, *16* (3), 252–261. <https://doi.org/10.1097/PHH.0b013e3181ce594e>.

- (169) Cooley, P.; Lee, B. Y.; Brown, S.; Cajka, J.; Chasteen, B.; Ganapathi, L.; Stark, J. H.; Wheaton, W. D.; Wagener, D. K.; Burke, D. S. Protecting Health Care Workers: A Pandemic Simulation Based on Allegheny County. *Influenza Other Respir. Viruses* **2010**, *4* (2), 61–72. <https://doi.org/10.1111/j.1750-2659.2009.00122.x>.
- (170) *Modeling the opioid crisis: How data reveal patterns in opioid misuse and overdose by location* | RTI. <https://www.rti.org/insights/modeling-opioid-crisis-how-data-reveal-patterns-opioid-misuse-and-overdose-location> (accessed 2024-09-07).
- (171) Bates, S.; Leonenko, V.; Rineer, J.; Bobashev, G. Using Synthetic Populations to Understand Geospatial Patterns in Opioid Related Overdose and Predicted Opioid Misuse. *Comput. Math. Organ. Theory* **2019**, *25* (1), 36–47. <https://doi.org/10.1007/s10588-018-09281-2>.
- (172) RTI International named an Obesity Data Challenge winner | RTI. <https://www.rti.org/news/rti-international-named-obesity-data-challenge-winner> (accessed 2024-09-07).
- (173) Appolon, G.; Tang, S.; Gabriel, N.; Morales, J.; Berenbrok, L. A.; LaCroix, A. Z.; Guo, J.; Mathis, W. S.; Hernandez, I. Racial and Ethnic Inequities in Spatial Access to Pharmacies: A Geographic Information System Analysis. *J. Am. Pharm. Assoc.* **2024**, *64* (4), 102131. <https://doi.org/10.1016/j.japh.2024.102131>.
- (174) Appolon, G.; Tang, S.; Gabriel, N.; Morales, J.; Berenbrok, L. A.; Guo, J.; Hernandez, I. Association Between Redlining and Spatial Access to Pharmacies. *JAMA Netw. Open* **2023**, *6* (8), e2327315. <https://doi.org/10.1001/jamanetworkopen.2023.27315>.
- (175) Haynes, D.; Hughes, K. D.; Borrero, M.; Haas, M.; Roach, L.; Blaes, A. Estimating Uninsured and Underinsured Women Eligible for Minnesota’s Breast Cancer Screening Program. *Cancer Causes Control* **2024**, *35* (2), 359–366. <https://doi.org/10.1007/s10552-023-01792-x>.
- (176) Bartsch, S. M.; Taitel, M. S.; DePasse, J. V.; Cox, S. N.; Smith-Ray, R. L.; Wedlock, P.; Singh, T. G.; Carr, S.; Siegmund, S. S.; Lee, B. Y. Epidemiologic and Economic Impact of Pharmacies as Vaccination Locations during an Influenza Epidemic. *Vaccine* **2018**, *36* (46), 7054–7063. <https://doi.org/10.1016/j.vaccine.2018.09.040>.
- (177) Stone, B. Jr.; Gronlund, C. J.; Mallen, E.; Hondula, D.; O’Neill, M. S.; Rajput, M.; Grijalva, S.; Lanza, K.; Harlan, S.; Larsen, L.; Augenbroe, G.; Krayenhoff, E. S.; Broadbent, A.; Georgescu, M. How Blackouts during Heat Waves Amplify Mortality and Morbidity Risk. *Environ. Sci. Technol.* **2023**, *57* (22), 8245–8255. <https://doi.org/10.1021/acs.est.2c09588>.
- (178) Kim, S. J.; Kery, C.; An, J.; Rineer, J.; Bobashev, G.; Matthews, A. K. Racial/Ethnic Disparities in Exposure to Neighborhood Violence and Lung Cancer Risk in Chicago. *Soc. Sci. Med.* **2024**, *340*, 116448. <https://doi.org/10.1016/j.socscimed.2023.116448>.
- (179) Leonenko, V. N.; Kovalchuk, S. V. Analyzing the Spatial Distribution of Individuals Predisposed to Arterial Hypertension in Saint Petersburg Using Synthetic Populations. *ITM Web Conf.* **2020**, *31*, 03002. <https://doi.org/10.1051/itmconf/20203103002>.
- (180) *Neighborhood Map of U.S. Obesity*. <https://synthpopviewer.rti.org/obesity/> (accessed 2024-09-07).
- (181) Rose, A. N.; Nagle, N. N. Validation of Spatiodemographic Estimates Produced through Data Fusion of Small Area Census Records and Household Microdata. *Comput. Environ. Urban Syst.* **2017**, *63*, 38–49. <https://doi.org/10.1016/j.compenurbsys.2016.07.006>.
- (182) Krieger, N. The US Census and the People’s Health: Public Health Engagement From Enslavement and “Indians Not Taxed” to Census Tracts and Health Equity (1790–2018). *Am. J. Public Health* **2019**, *109* (8), 1092–1100. <https://doi.org/10.2105/AJPH.2019.305017>.
- (183) *Counting America’s First Peoples - Carolyn A. Liebler, 2018*. <https://journals.sagepub.com/doi/abs/10.1177/0002716218766276> (accessed 2024-09-12).

- (184) Jim, M. A.; Arias, E.; Seneca, D. S.; Hoopes, M. J.; Jim, C. C.; Johnson, N. J.; Wiggins, C. L. Racial Misclassification of American Indians and Alaska Natives by Indian Health Service Contract Health Service Delivery Area. *Am. J. Public Health* **2014**, *104* (S3), S295–S302. <https://doi.org/10.2105/AJPH.2014.301933>.
- (185) Draghi, B.; Wang, Z.; Myles, P.; Tucker, A. Identifying and Handling Data Bias within Primary Healthcare Data Using Synthetic Data Generators. *Heliyon* **2024**, *10* (2), e24164. <https://doi.org/10.1016/j.heliyon.2024.e24164>.
- (186) Boyd, A. D.; Gonzalez-Guarda, R.; Lawrence, K.; Patil, C. L.; Ezenwa, M. O.; O'Brien, E. C.; Paek, H.; Braciszewski, J. M.; Adeyemi, O.; Cuthel, A. M.; Darby, J. E.; Zigler, C. K.; Ho, P. M.; Faurot, K. R.; Staman, K. L.; Leigh, J. W.; Dailey, D. L.; Cheville, A.; Del Fiol, G.; Knisely, M. R.; Grudzen, C. R.; Marsolo, K.; Richesson, R. L.; Schlaeger, J. M. Potential Bias and Lack of Generalizability in Electronic Health Record Data: Reflections on Health Equity from the National Institutes of Health Pragmatic Trials Collaboratory. *J. Am. Med. Inform. Assoc.* **2023**, *30* (9), 1561–1566. <https://doi.org/10.1093/jamia/ocad115>.
- (187) Kitman, J. L. The Secret History of Lead. March 2, 2000. <https://www.thenation.com/article/archive/secret-history-lead/> (accessed 2024-10-07).
- (188) Dignam, T.; Kaufmann, R. B.; LeStourgeon, L.; Brown, M. J. Control of Lead Sources in the United States, 1970-2017: Public Health Progress and Current Challenges to Eliminating Lead Exposure. *J. Public Health Manag. Pract. JPHMP* **2019**, *25* (Suppl 1 LEAD POISONING PREVENTION), S13–S22. <https://doi.org/10.1097/PHH.0000000000000889>.
- (189) Bellinger, D. C.; Bellinger, A. M. Childhood Lead Poisoning: The Torturous Path from Science to Policy. *J. Clin. Invest.* **2006**, *116* (4), 853–857. <https://doi.org/10.1172/JCI28232>.
- (190) Toxicological Profile for Lead. <https://doi.org/10.15620/cdc.95222>.
- (191) COUNCIL ON ENVIRONMENTAL HEALTH; Lanphear, B. P.; Lowry, J. A.; Ahdoot, S.; Baum, C. R.; Bernstein, A. S.; Bole, A.; Brumberg, H. L.; Campbell, C. C.; Lanphear, B. P.; Pacheco, S. E.; Spanier, A. J.; Trasande, L. Prevention of Childhood Lead Toxicity. *Pediatrics* **2016**, *138* (1), e20161493. <https://doi.org/10.1542/peds.2016-1493>.
- (192) Hornung, R. W.; Lanphear, B. P.; Dietrich, K. N. Age of Greatest Susceptibility to Childhood Lead Exposure: A New Statistical Approach. *Environ. Health Perspect.* **2009**, *117* (8), 1309–1312. <https://doi.org/10.1289/ehp.0800426>.
- (193) Abelsohn, A. R.; Sanborn, M. Lead and Children. *Can. Fam. Physician* **2010**, *56* (6), 531–535.
- (194) Larsen, B.; Sánchez-Triana, E. Global Health Burden and Cost of Lead Exposure in Children and Adults: A Health Impact and Economic Modelling Analysis. *Lancet Planet. Health* **2023**, *7* (10), e831–e840. [https://doi.org/10.1016/S2542-5196\(23\)00166-3](https://doi.org/10.1016/S2542-5196(23)00166-3).
- (195) Gould, E. Childhood Lead Poisoning: Conservative Estimates of the Social and Economic Benefits of Lead Hazard Control. *Environ. Health Perspect.* **2009**, *117* (7), 1162–1167. <https://doi.org/10.1289/ehp.0800408>.
- (196) Grosse, S. D.; Matte, T. D.; Schwartz, J.; Jackson, R. J. *Economic gains resulting from the reduction in children's exposure to lead in the United States*. <https://ehp.niehs.nih.gov/doi/epdf/10.1289/ehp.02110563> (accessed 2024-10-07).
- (197) Egan, K. B.; Cornwell, C. R.; Courtney, J. G.; Ettinger, A. S. Blood Lead Levels in U.S. Children Ages 1–11 Years, 1976–2016. *Environ. Health Perspect.* **2021**, *129* (3), 037003. <https://doi.org/10.1289/EHP7932>.
- (198) Centers for Disease Control and Prevention (CDC). Blood Lead Levels in Children Aged 1-5 Years - United States, 1999-2010. *MMWR Morb. Mortal. Wkly. Rep.* **2013**, *62* (13), 245–248.
- (199) Teye, S. O.; Yanosky, J. D.; Cuffee, Y.; Weng, X.; Luquis, R.; Farace, E.; Wang, L. Exploring Persistent Racial/Ethnic Disparities in Lead Exposure among American Children Aged 1–5

- Years: Results from NHANES 1999–2016. *Int. Arch. Occup. Environ. Health* **2021**, *94* (4), 723–730. <https://doi.org/10.1007/s00420-020-01616-4>.
- (200) Lanphear, B. P.; Weitzman, M.; Eberly, S. Racial Differences in Urban Children’s Environmental Exposures to Lead. *Am. J. Public Health* **1996**, *86* (10), 1460–1463. <https://doi.org/10.2105/AJPH.86.10.1460>.
- (201) Hauptman, M.; Rogers, M. L.; Scarpaci, M.; Morin, B.; Vivier, P. M. Neighborhood Disparities and the Burden of Lead Poisoning. *Pediatr. Res.* **2023**, *94* (2), 826–836. <https://doi.org/10.1038/s41390-023-02476-7>.
- (202) Ruckart, P. Z.; Jones, R. L.; Courtney, J. G.; LeBlanc, T. T.; Jackson, W.; Karwowski, M. P.; Cheng, P.-Y.; Allwood, P.; Svendsen, E. R.; Breysse, P. N. Update of the Blood Lead Reference Value — United States, 2021. *Morb. Mortal. Wkly. Rep.* **2021**, *70* (43), 1509–1512. <https://doi.org/10.15585/mmwr.mm7043a4>.
- (203) Farr, N.; Dolbeare, C. N. Childhood Lead Poisoning: Solving a Health and Housing Problem.
- (204) Noble, D. Blood Lead Measurement Takes the Direct Approach. *Anal. Chem.* **1993**, *65* (5), 265A–267A. <https://doi.org/10.1021/ac00053a002>.
- (205) Children at Risk: Gaps in State Lead Screening Policies. https://toxicfreefuture.org/wp-content/uploads/2017/01/saferchemicals.org_children-at-risk-report.pdf (accessed 2024-10-08).
- (206) *State Lead Screening Policies and Reporting Requirements*. Toxic-Free Future. <https://toxicfreefuture.org/research/children-at-risk/state-lead-screening-policies/> (accessed 2024-10-07).
- (207) State Health Department Policies for Children with Elevated Blood Lead Levels. <https://nchh.org/resource-library/state-health-department-policies-for-children-with-elevated-blood-lead-levels.pdf> (accessed 2024-10-07).
- (208) Hauptman, M.; Niles, J. K.; Gudin, J.; Kaufman, H. W. Individual- and Community-Level Factors Associated With Detectable and Elevated Blood Lead Levels in US Children: Results From a National Clinical Laboratory. *JAMA Pediatr.* **2021**, *175* (12), 1252–1260. <https://doi.org/10.1001/jamapediatrics.2021.3518>.
- (209) Pell, M. B.; Schneyer, J. *Reuters finds lead levels higher than Flint’s in thousands of locales*. Reuters. <http://www.reuters.com/investigates/special-report/usa-lead-testing/> (accessed 2024-10-08).
- (210) Chen, Y.-H.; Ma, Z.-Q.; Watkins, S. M. Effects of Individual and Neighborhood Characteristics on Childhood Blood Lead Testing and Elevated Blood Lead Levels, A Pennsylvania Birth Cohort Analysis. *J. Prim. Care Community Health* **2021**, *12*, 21501327211017780. <https://doi.org/10.1177/21501327211017780>.
- (211) Kamai, E. M.; Daniels, J. L.; Delamater, P. L.; Lanphear, B. P.; MacDonald Gibson, J.; Richardson, D. B. Patterns of Children’s Blood Lead Screening and Blood Lead Levels in North Carolina, 2011–2018—Who Is Tested, Who Is Missed? *Environ. Health Perspect.* **2022**, *130* (6), 067002. <https://doi.org/10.1289/EHP10335>.
- (212) Abbasi, A.; DiTraglia, F. J.; Gazze, L.; Pals, B. Hidden Hazards and Screening Policy: Predicting Undetected Lead Exposure in Illinois. *J. Health Econ.* **2023**, *90*, 102783. <https://doi.org/10.1016/j.jhealeco.2023.102783>.
- (213) US EPA, O. *Federal Action Plan to Reduce Childhood Lead Exposure*. <https://www.epa.gov/lead/federal-action-plan-reduce-childhood-lead-exposure> (accessed 2024-10-08).
- (214) US EPA, O. *Data Mapping to Identify High Lead Exposure Risk Locations in the U.S.* <https://www.epa.gov/lead/mapping> (accessed 2024-10-08).

- (215) *Targeting Coordinated Federal Efforts to Address Persistent Hazardous Exposures to Lead*. <https://doi.org/10.2105/AJPH.2022.306972>.
- (216) Zartarian, V.; Poulakos, A.; Garrison, V. H.; Spalt, N.; Tornero-Velez, R.; Xue, J.; Egan, K.; Courtney, J. Lead Data Mapping to Prioritize US Locations for Whole-of-Government Exposure Prevention Efforts: State of the Science, Federal Collaborations, and Remaining Challenges. *Am. J. Public Health* **2022**, *112* (Suppl 7), S658–S669. <https://doi.org/10.2105/AJPH.2022.307051>.
- (217) Egan, K.; Courtney, J. CDC-ATSDR Lead Exposure Risk Index (LERI).
- (218) Frostenson, S.; April 6, S. K. on; 2016; ET, 8:50 a m. *Where is the lead exposure risk in your community?*. Vox.com. <http://www.vox.com/a/lead-exposure-risk-map> (accessed 2024-10-07).
- (219) Schultz, B. D.; Morara, M.; Buxton, B. E.; Weintraub, M. Predicting Blood-Lead Levels Among U.S. Children at the Census Tract Level. *Environ. Justice* **2017**, *10* (5), 129–136. <https://doi.org/10.1089/env.2017.0005>.
- (220) US EPA, O. *Final Report for the Pilot Study of Targeting Elevated Blood-Lead Levels in Children*, EPA 260-R-08-003, October 3, 2007. <https://www.epa.gov/lead/final-report-pilot-study-targeting-elevated-blood-lead-levels-children-epa-260-r-08-003> (accessed 2024-10-23).
- (221) Xue, J.; Zartarian, V.; Tornero-Velez, R.; Stanek, L. W.; Poulakos, A.; Walts, A.; Triantafillou, K.; Suero, M.; Grokhowsky, N. A Generalizable Evaluated Approach, Applying Advanced Geospatial Statistical Methods, to Identify High Lead Exposure Locations at Census Tract Scale: Michigan Case Study. *Environ. Health Perspect.* **2022**, *130* (7), 077004. <https://doi.org/10.1289/EHP9705>.
- (222) Stanek, L. W.; Xue, J.; Zartarian, V. G.; Poulakos, A. G.; Tornero-Velez, R.; Snyder, E. G.; Walts, A.; Triantafillou, K. Identification of High Lead Exposure Locations in Ohio at the Census Tract Scale Using a Generalizable Geospatial Hotspot Approach. *J. Expo. Sci. Environ. Epidemiol.* **2024**, *34* (4), 718–726. <https://doi.org/10.1038/s41370-024-00666-x>.
- (223) Zartarian, V. G.; Xue, J.; Poulakos, A. G.; Tornero-Velez, R.; Stanek, L. W.; Snyder, E.; Helms Garrison, V.; Egan, K.; Courtney, J. G. A U.S. Lead Exposure Hotspots Analysis. *Environ. Sci. Technol.* **2024**, *58* (7), 3311–3321. <https://doi.org/10.1021/acs.est.3c07881>.
- (224) Wheeler, D. C.; Boyle, J.; Raman, S.; Nelson, E. J. Modeling Elevated Blood Lead Level Risk across the United States. *Sci. Total Environ.* **2021**, *769*, 145237. <https://doi.org/10.1016/j.scitotenv.2021.145237>.
- (225) Lobo, G. P.; Kalyan, B.; Gadgil, A. J. Predicting Childhood Lead Exposure at an Aggregated Level Using Machine Learning. *Int. J. Hyg. Environ. Health* **2021**, *238*, 113862. <https://doi.org/10.1016/j.ijheh.2021.113862>.
- (226) Potash, E.; Ghani, R.; Walsh, J.; Jorgensen, E.; Lohff, C.; Prachand, N.; Mansour, R. Validation of a Machine Learning Model to Predict Childhood Lead Poisoning. *JAMA Netw. Open* **2020**, *3* (9), e2012734. <https://doi.org/10.1001/jamanetworkopen.2020.12734>.
- (227) Frndak, S.; Yan, F.; Edelson, M.; Immergluck, L. C.; Kordas, K.; Idris, M. Y.; Dickinson-Copeland, C. M. Predicting Low-Level Childhood Lead Exposure in Metro Atlanta Using Ensemble Machine Learning of High-Resolution Raster Cells. *Int. J. Environ. Res. Public Health* **2023**, *20* (5), 4477. <https://doi.org/10.3390/ijerph20054477>.
- (228) Lanphear, B. P.; Hornung, R.; Ho, M. Screening Housing to Prevent Lead Toxicity in Children. *Public Health Rep.* **2005**, *120* (3), 305–310.
- (229) Turrini, G.; Branham, D. K.; Chen, L.; Conmy, A. B.; Chappel, A. R.; Lew, N. D.; Sommers, B. D. Access to Affordable Care in Rural America: Current Trends and Key Challenges. *Assist. Sec. Plan. Eval.* **2021**.

- (230) *Rural Americans Struggle with Medical Bills and Health Care Affordability*. <https://doi.org/10.26099/pq3a-k123>.
- (231) Davis, J. C.; Rupasingha, A.; Cromartie, J.; Sanders, A. *Rural America at a Glance: 2022 Edition*. <http://www.ers.usda.gov/publications/pub-details/?pubid=105154> (accessed 2024-10-08).
- (232) Eh, N.; Wc, B.; I, H.-P.; Da, N. Rural-Urban Blood Lead Differences in North Carolina Children. *Pediatrics* **1994**, *94* (1).
- (233) Housing In Rural America. https://ruralhome.org/wp-content/uploads/storage/documents/ts2010/ts-report/ts10_rural_housing.pdf (accessed 2024-10-08).
- (234) Homeownership in Rural America. https://ruralhome.org/wp-content/uploads/2021/05/Homeownership_in_Rura_America_web.pdf (accessed 2024-10-08).
- (235) Rural Communities and Healthy Housing Primer. <https://nchh.org/resource-library/rural-communities-and-healthy-housing-primer.pdf> (accessed 2024-10-08).
- (236) Yam, H. Rural/Urban Differences in Housing Quality and Adequacy: Findings from the American Housing Survey, 2019. **2019**.
- (237) Indian Lands of Federally Recognized Tribes. <https://www.bia.gov/sites/default/files/dup/assets/bia/ots/webteam/pdf/idc1-028635.pdf> (accessed 2024-10-08).
- (238) *Indian Reservations - Map, US & Definition*. HISTORY. <https://www.history.com/topics/native-american-history/indian-reservations> (accessed 2024-10-08).
- (239) Lofquist, D. A. Multigenerational Households: 2009–2011.
- (240) *Housing Needs of American Indians and Alaska Natives in Tribal Areas: A Report From the Assessment of American Indian, Alaska Native, and Native Hawaiian Housing Needs* | HUD USER. <https://www.huduser.gov/portal/publications/HNAIHousingNeeds.html> (accessed 2024-10-08).
- (241) Ranco, D. J.; O'Neill, C. A.; Donatuto, J.; Harper, B. L. Environmental Justice, American Indians and the Cultural Dilemma: Developing Environmental Management for Tribal Health and Well-Being. *Environ. Justice* **2011**, *4* (4), 221–230. <https://doi.org/10.1089/env.2010.0036>.
- (242) Hoover, E.; Cook, K.; Plain, R.; Sanchez, K.; Waghiyi, V.; Miller, P.; Dufault, R.; Sislin, C.; Carpenter, D. O. Indigenous Peoples of North America: Environmental Exposures and Reproductive Justice. *Environ. Health Perspect.* **2012**, *120* (12), 1645–1649. <https://doi.org/10.1289/ehp.1205422>.
- (243) Vickery, J.; Hunter, L. M. Native Americans: Where in Environmental Justice Research? *Soc. Nat. Resour.* **2016**, *29* (1), 36–52. <https://doi.org/10.1080/08941920.2015.1045644>.
- (244) Gochfeld, M.; Burger, J. Disproportionate Exposures in Environmental Justice and Other Populations: The Importance of Outliers. *Am. J. Public Health* **2011**, *101* (Suppl 1), S53–S63. <https://doi.org/10.2105/AJPH.2011.300121>.
- (245) Malcoe, L. H.; Lynch, R. A.; Keger, M. C.; Skaggs, V. J. Lead Sources, Behaviors, and Socioeconomic Factors in Relation to Blood Lead of Native American and White Children: A Community-Based Assessment of a Former Mining Area. *Environ. Health Perspect.* **2002**, *110* (Suppl 2), 221–231.
- (246) Bullard, R. D.; Clinton, W. J. Overcoming Racism in Environmental Decisionmaking. *Environment* **1994**, *36* (4), 10.
- (247) White, B. M.; Bonilha, H. S.; Ellis, C. Racial/Ethnic Differences in Childhood Blood Lead Levels Among Children <72 Months of Age in the United States: A Systematic Review of the

- Literature. *J. Racial Ethn. Health Disparities* **2016**, *3* (1), 145–153.
<https://doi.org/10.1007/s40615-015-0124-9>.
- (248) Leech, T. G. J.; Adams, E. A.; Weathers, T. D.; Staten, L. K.; Filippelli, G. M. Inequitable Chronic Lead Exposure: A Dual Legacy of Social and Environmental Injustice. *Fam. Community Health* **2016**, *39* (3), 151. <https://doi.org/10.1097/FCH.000000000000106>.
- (249) Lanphear, B. P.; Hornung, R.; Ho, M.; Howard, C. R.; Eberly, S.; Knauf, K. Environmental Lead Exposure during Early Childhood. *J. Pediatr.* **2002**, *140* (1), 40–47.
<https://doi.org/10.1067/mpd.2002.120513>.
- (250) Massey, D. S. The Legacy of the 1968 Fair Housing Act. *Sociol. Forum Randolph NJ* **2015**, *30* (Suppl 1), 571–588. <https://doi.org/10.1111/socf.12178>.
- (251) *Conceptualizing the New Urban Renewal: Comparing the Past to the Present - Derek S. Hyra, 2012.*
<https://journals.sagepub.com/doi/abs/10.1177/1078087411434905> (accessed 2024-10-08).
- (252) Rugh, J. S.; Massey, D. S. Racial Segregation and the American Foreclosure Crisis. *Am. Sociol. Rev.* **2010**, *75* (5), 629–651. <https://doi.org/10.1177/0003122410380868>.
- (253) Jacobs, D. E. Environmental Health Disparities in Housing. *Am. J. Public Health* **2011**, *101* (S1), S115–S122. <https://doi.org/10.2105/AJPH.2010.300058>.
- (254) Rosenbaum, E. Racial/Ethnic Differences in Home Ownership and Housing Quality, 1991*. *Soc. Probl.* **1996**, *43* (4), 403–426. <https://doi.org/10.2307/3096952>.
- (255) Haurin, D. R.; Herbert, C. E.; Rosenthal, S. S. Homeownership Gaps Among Low-Income and Minority Households. *Cityscape* **2007**, *9* (2), 5–51.
- (256) Neuwirth, L. S. Resurgent Lead Poisoning and Renewed Public Attention towards Environmental Social Justice Issues: A Review of Current Efforts and Call to Revitalize Primary and Secondary Lead Poisoning Prevention for Pregnant Women, Lactating Mothers, and Children within the U.S. *Int. J. Occup. Environ. Health* **2018**, *24* (3–4), 86–100.
<https://doi.org/10.1080/10773525.2018.1507291>.
- (257) Garrison, V. E. H.; Ashley, P. J. Identifying Jurisdictions at Risk of Containing Housing Units With Deteriorated Paint: Results and Targeting Implications for the US Department of Housing and Urban Development. *J. Public Health Manag. Pract.* **2020**, *27* (6), 546.
<https://doi.org/10.1097/PHH.0000000000001191>.
- (258) *Finding Lead Water Pipes: New NRDC Map Shows the Hot Spots in Every State.*
<https://www.nrdc.org/resources/lead-pipes-are-widespread-and-used-every-state> (accessed 2024-10-24).
- (259) Liu, G.; Chen, H.; Yuan, Y.; Song, C. Indoor Thermal Environment and Human Health: A Systematic Review. *Renew. Sustain. Energy Rev.* **2024**, *191*, 114164.
<https://doi.org/10.1016/j.rser.2023.114164>.
- (260) Cedeño Laurent, J. G.; Williams, A.; Oulhote, A.; Zanolletti, A.; Allen, J. G.; Spengler, J. D. *Reduced cognitive function during a heat wave among residents of non-air-conditioned buildings: An observational study of young adults in the summer of 2016 | PLOS Medicine.*
<https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002605> (accessed 2024-11-06).
- (261) Braga, A. L. F.; Zanolletti, A.; Schwartz, J. The Effect of Weather on Respiratory and Cardiovascular Deaths in 12 U.S. Cities. *Environ. Health Perspect.* **2002**, *110* (9), 859–863.
<https://doi.org/10.1289/ehp.02110859>.
- (262) Ebi, K. L.; Capon, A.; Berry, P.; Broderick, C.; Dear, R. de; Havenith, G.; Honda, Y.; Kovats, R. S.; Ma, W.; Malik, A.; Morris, N. B.; Nybo, L.; Seneviratne, S. I.; Vanos, J.; Jay, O. Hot Weather and Heat Extremes: Health Risks. *The Lancet* **2021**, *398* (10301), 698–708.
[https://doi.org/10.1016/S0140-6736\(21\)01208-3](https://doi.org/10.1016/S0140-6736(21)01208-3).

- (263) Luber, G.; McGeehin, M. Climate Change and Extreme Heat Events. *Am. J. Prev. Med.* **2008**, *35* (5), 429–435. <https://doi.org/10.1016/j.amepre.2008.08.021>.
- (264) Ellis, F. P. Mortality from Heat Illness and Heat-Aggravated Illness in the United States. *Environ. Res.* **1972**, *5* (1), 1–58. [https://doi.org/10.1016/0013-9351\(72\)90019-9](https://doi.org/10.1016/0013-9351(72)90019-9).
- (265) Anderson, B. G.; Bell, M. L. Weather-Related Mortality: How Heat, Cold, and Heat Waves Affect Mortality in the United States. *Epidemiol. Camb. Mass* **2009**, *20* (2), 205. <https://doi.org/10.1097/EDE.0b013e318190ee08>.
- (266) *The Health Care Costs of Extreme Heat*. Center for American Progress. <https://www.americanprogress.org/article/the-health-care-costs-of-extreme-heat/> (accessed 2024-11-06).
- (267) Atlantic Council of the United States. *Extreme Heat: The Economic and Social Consequences for the United States*; Atlantic Council: Washington, DC, 2021.
- (268) Committee, U. S. J. E. *The Mounting Costs of Extreme Heat - The Mounting Costs of Extreme Heat - United States Joint Economic Committee*. <https://www.jec.senate.gov/public/index.cfm/democrats/2023/8/the-mounting-costs-of-extreme-heat> (accessed 2024-11-06).
- (269) Klepeis, N. E.; Nelson, W. C.; Ott, W. R.; Robinson, J. P.; Tsang, A. M.; Switzer, P.; Behar, J. V.; Hern, S. C.; Engelmann, W. H. The National Human Activity Pattern Survey (NHAPS): A Resource for Assessing Exposure to Environmental Pollutants. *J. Expo. Sci. Environ. Epidemiol.* **2001**, *11* (3), 231–252. <https://doi.org/10.1038/sj.jea.7500165>.
- (270) Barreca, A.; Clay, K.; Deschenes, O.; Greenstone, M.; Shapiro, J. S. Adapting to Climate Change: The Remarkable Decline in the US Temperature-Mortality Relationship over the Twentieth Century. *J. Polit. Econ.* **2016**, *124* (1), 105–159. <https://doi.org/10.1086/684582>.
- (271) Hondula, D. M.; Balling, R. C.; Vanos, J. K.; Georgescu, M. Rising Temperatures, Human Health, and the Role of Adaptation. *Curr. Clim. Change Rep.* **2015**, *1* (3), 144–154. <https://doi.org/10.1007/s40641-015-0016-4>.
- (272) Sera, F.; Hashizume, M.; Honda, Y.; Lavigne, E.; Schwartz, J.; Zanobetti, A.; Tobias, A.; Iñiguez, C.; Vicedo-Cabrera, A. M.; Blangiardo, M.; Armstrong, B.; Gasparrini, A. Air Conditioning and Heat-Related Mortality: A Multi-Country Longitudinal Study. *Epidemiol. Camb. Mass* **2020**, *31* (6), 779–787. <https://doi.org/10.1097/EDE.0000000000001241>.
- (273) Kinney, P. L. Temporal Trends in Heat-Related Mortality: Implications for Future Projections. *Atmosphere* **2018**, *9* (10), 409. <https://doi.org/10.3390/atmos9100409>.
- (274) *Nearly 90% of U.S. households used air conditioning in 2020 - U.S. Energy Information Administration (EIA)*. <https://www.eia.gov/todayinenergy/detail.php?id=52558> (accessed 2024-11-06).
- (275) Lundgren-Kownacki, K.; Hornyanszky, E. D.; Chu, T. A.; Olsson, J. A.; Becker, P. Challenges of Using Air Conditioning in an Increasingly Hot Climate. *Int. J. Biometeorol.* **2018**, *62* (3), 401–412. <https://doi.org/10.1007/s00484-017-1493-z>.
- (276) Randazzo, T.; De Cian, E.; Mistry, M. N. Air Conditioning and Electricity Expenditure: The Role of Climate in Temperate Countries. *Econ. Model.* **2020**, *90*, 273–287. <https://doi.org/10.1016/j.econmod.2020.05.001>.
- (277) Batlle, C.; Heller, P.; Knittel, C.; Schittekatte, T. US Federal Resource Allocations Are Inconsistent with Concentrations of Energy Poverty. *Sci. Adv.* **2024**, *10* (41), eadp8183. <https://doi.org/10.1126/sciadv.adp8183>.
- (278) Hernández, D.; Laird, J. Surviving a Shut-Off: U.S. Households at Greatest Risk of Utility Disconnections and How They Cope. *Am. Behav. Sci.* **2022**, *66* (7), 856–880. <https://doi.org/10.1177/00027642211013401>.

- (279) RECS: One in three U.S. households faced challenges in paying energy bills in 2015. <https://www.eia.gov/consumption/residential/reports/2015/energybills/> (accessed 2024-11-06).
- (280) Obringer, R.; Nateghi, R.; Maia-Silva, D.; Mukherjee, S.; Cr, V.; McRoberts, D. B.; Kumar, R. Implications of Increasing Household Air Conditioning Use Across the United States Under a Warming Climate. *Earths Future* **2022**, *10* (1), e2021EF002434. <https://doi.org/10.1029/2021EF002434>.
- (281) Jones, A.; Nock, D.; Samaras, C.; Qiu, Y. (Lucy); Xing, B. Climate Change Impacts on Future Residential Electricity Consumption and Energy Burden: A Case Study in Phoenix, Arizona. *Energy Policy* **2023**, *183*, 113811. <https://doi.org/10.1016/j.enpol.2023.113811>.
- (282) Bednar, D. J.; Reames, T. G. Recognition of and Response to Energy Poverty in the United States. *Nat. Energy* **2020**, *5* (6), 432–439. <https://doi.org/10.1038/s41560-020-0582-0>.
- (283) Falchetta, G.; Cian, E. D.; Pavanello, F.; Wing, I. S. Inequalities in Global Residential Cooling Energy Use to 2050. *Nat. Commun.* **2024**, *15* (1), 7874. <https://doi.org/10.1038/s41467-024-52028-8>.
- (284) Ortiz, L.; Gamarro, H.; Gonzalez, J. E.; McPhearson, T. Energy Burden and Air Conditioning Adoption in New York City under a Warming Climate. *Sustain. Cities Soc.* **2022**, *76*, 103465. <https://doi.org/10.1016/j.scs.2021.103465>.
- (285) Hoffman, J. S.; Shandas, V.; Pendleton, N. The Effects of Historical Housing Policies on Resident Exposure to Intra-Urban Heat: A Study of 108 US Urban Areas. *Climate* **2020**, *8* (1), 12. <https://doi.org/10.3390/cli8010012>.
- (286) Wilson, B. Urban Heat Management and the Legacy of Redlining. *J. Am. Plann. Assoc.* **2020**, *86* (4), 443–457. <https://doi.org/10.1080/01944363.2020.1759127>.
- (287) Benz, S. A.; Burney, J. A. Widespread Race and Class Disparities in Surface Urban Heat Extremes Across the United States. *Earths Future* **2021**, *9* (7), e2021EF002016. <https://doi.org/10.1029/2021EF002016>.
- (288) Mitchell, B. C.; Chakraborty, J. Landscapes of Thermal Inequity: Disproportionate Exposure to Urban Heat in the Three Largest US Cities. *Environ. Res. Lett.* **2015**, *10* (11), 115005. <https://doi.org/10.1088/1748-9326/10/11/115005>.
- (289) Gronlund, C. J. Racial and Socioeconomic Disparities in Heat-Related Health Effects and Their Mechanisms: A Review. *Curr. Epidemiol. Rep.* **2014**, *1* (3), 165. <https://doi.org/10.1007/s40471-014-0014-4>.
- (290) O'Neill, M. S.; Zanobetti, A.; Schwartz, J. Disparities by Race in Heat-Related Mortality in Four US Cities: The Role of Air Conditioning Prevalence. *J. Urban Health* **2005**, *82* (2), 191–197. <https://doi.org/10.1093/jurban/jti043>.
- (291) Romitti, Y.; Wing, I. S.; Spangler, K. R.; Wellenius, G. A. *Inequality in the availability of residential air conditioning across 115 US metropolitan areas | PNAS Nexus | Oxford Academic*. <https://academic.oup.com/pnasnexus/article/1/4/pgac210/6713629> (accessed 2024-11-06).
- (292) Jessel, S.; Sawyer, S.; Hernández, D. Energy, Poverty, and Health in Climate Change: A Comprehensive Review of an Emerging Literature. *Front. Public Health* **2019**, *7*, 357. <https://doi.org/10.3389/fpubh.2019.00357>.
- (293) Goldstein, B.; Reames, T. G.; Newell, J. P. Racial Inequity in Household Energy Efficiency and Carbon Emissions in the United States: An Emissions Paradox. *Energy Res. Soc. Sci.* **2022**, *84*, 102365. <https://doi.org/10.1016/j.erss.2021.102365>.
- (294) Wang, Q.; Kwan, M.-P.; Fan, J.; Lin, J. Racial Disparities in Energy Poverty in the United States. *Renew. Sustain. Energy Rev.* **2021**, *137*, 110620. <https://doi.org/10.1016/j.rser.2020.110620>.

- (295) Romanello, M.; McGushin, A.; Napoli, C. D.; Drummond, P.; Hughes, N.; Jamart, L.; Kennard, H.; Lampard, P.; Rodriguez, B. S.; Arnell, N.; Ayeb-Karlsson, S.; Belesova, K.; Cai, W.; Campbell-Lendrum, D.; Capstick, S.; Chambers, J.; Chu, L.; Ciampi, L.; Dalin, C.; Dasandi, N.; Dasgupta, S.; Davies, M.; Dominguez-Salas, P.; Dubrow, R.; Ebi, K. L.; Eckelman, M.; Ekins, P.; Escobar, L. E.; Georgeson, L.; Grace, D.; Graham, H.; Gunther, S. H.; Hartinger, S.; He, K.; Heaviside, C.; Hess, J.; Hsu, S.-C.; Jankin, S.; Jimenez, M. P.; Kelman, I.; Kieseewetter, G.; Kinney, P. L.; Kjellstrom, T.; Kniveton, D.; Lee, J. K. W.; Lemke, B.; Liu, Y.; Liu, Z.; Lott, M.; Lowe, R.; Martinez-Urtaza, J.; Maslin, M.; McAllister, L.; McMichael, C.; Mi, Z.; Milner, J.; Minor, K.; Mohajeri, N.; Moradi-Lakeh, M.; Morrissey, K.; Munzert, S.; Murray, K. A.; Neville, T.; Nilsson, M.; Obradovich, N.; Sewe, M. O.; Oreszczyn, T.; Otto, M.; Owfi, F.; Pearman, O.; Pencheon, D.; Rabbaniha, M.; Robinson, E.; Rocklöv, J.; Salas, R. N.; Semenza, J. C.; Sherman, J.; Shi, L.; Springmann, M.; Tabatabaei, M.; Taylor, J.; Trinanes, J.; Shumake-Guillemot, J.; Vu, B.; Wagner, F.; Wilkinson, P.; Winning, M.; Yglesias, M.; Zhang, S.; Gong, P.; Montgomery, H.; Costello, A.; Hamilton, I. The 2021 Report of the Lancet Countdown on Health and Climate Change: Code Red for a Healthy Future. *The Lancet* **2021**, *398* (10311), 1619–1662. [https://doi.org/10.1016/S0140-6736\(21\)01787-6](https://doi.org/10.1016/S0140-6736(21)01787-6).
- (296) Ramesh, T.; Wozniak, G. D.; Yu, H. County-Level Disparities in Heat-Related Emergencies. *JAMA Netw. Open* **2024**, *7* (3), e242845. <https://doi.org/10.1001/jamanetworkopen.2024.2845>.
- (297) Berberian, A. G.; Gonzalez, D. J. X.; Cushing, L. J. Racial Disparities in Climate Change-Related Health Effects in the United States. *Curr. Environ. Health Rep.* **2022**, *9* (3), 451–464. <https://doi.org/10.1007/s40572-022-00360-w>.
- (298) Lee, E. K.; Donley, G.; Ciesielski, T. H.; Gill, I.; Yamoah, O.; Roche, A.; Martinez, R.; Freedman, D. A. Health Outcomes in Redlined versus Non-Redlined Neighborhoods: A Systematic Review and Meta-Analysis. *Soc. Sci. Med.* **2022**, *294*, 114696. <https://doi.org/10.1016/j.socscimed.2021.114696>.
- (299) Basu, R. High Ambient Temperature and Mortality: A Review of Epidemiologic Studies from 2001 to 2008. *Environ. Health* **2009**, *8* (1), 40. <https://doi.org/10.1186/1476-069X-8-40>.
- (300) Davis, L.; Gertler, P.; Jarvis, S.; Wolfram, C. Air Conditioning and Global Inequality. *Glob. Environ. Change* **2021**, *69*, 102299. <https://doi.org/10.1016/j.gloenvcha.2021.102299>.
- (301) Reid, C. E.; O'Neill, M. S.; Gronlund, C. J.; Brines, S. J.; Brown, D. G.; Diez-Roux, A. V.; Schwartz, J. Mapping Community Determinants of Heat Vulnerability. *Environ. Health Perspect.* **2009**, *117* (11), 1730–1736. <https://doi.org/10.1289/ehp.0900683>.
- (302) Bao, J.; Li, X.; Yu, C. The Construction and Validation of the Heat Vulnerability Index, a Review. *Int. J. Environ. Res. Public Health* **2015**, *12* (7), 7220. <https://doi.org/10.3390/ijerph120707220>.
- (303) Ahn, Y.; Uejio, C. K. Modeling Air Conditioning Ownership and Availability. *Urban Clim.* **2022**, *46*, 101322. <https://doi.org/10.1016/j.uclim.2022.101322>.
- (304) Ahn, Y.; Uejio, C. K.; Wong, S.; Powell, E.; Holmes, T. Spatial Disparities in Air Conditioning Ownership in Florida, United States.
- (305) Gronlund, C. J.; Berrocal, V. J. Modeling and Comparing Central and Room Air Conditioning Ownership and Cold-Season in-Home Thermal Comfort Using the American Housing Survey. *J. Expo. Sci. Environ. Epidemiol.* **2020**, *30* (5), 814–823. <https://doi.org/10.1038/s41370-020-0220-8>.
- (306) Guirguis, K.; Basu, R.; Al-Delaimy, W. K.; Benmarhnia, T.; Clemesha, R. E. S.; Corcos, I.; Guzman-Morales, J.; Hailey, B.; Small, I.; Tardy, A.; Vashishtha, D.; Zivin, J. G.; Gershunov,

- A. Heat, Disparities, and Health Outcomes in San Diego County's Diverse Climate Zones. *GeoHealth* **2018**, 2 (7), 212–223. <https://doi.org/10.1029/2017GH000127>.
- (307) Chen, M.; Sanders, K. T.; Ban-Weiss, G. A. A New Method Utilizing Smart Meter Data for Identifying the Existence of Air Conditioning in Residential Homes. *Environ. Res. Lett.* **2019**, 14 (9), 094004. <https://doi.org/10.1088/1748-9326/ab35a8>.
- (308) Fraser, A. M.; Chester, M. V.; Eisenman, D.; Hondula, D. M.; Pincetl, S. S.; English, P.; Bondank, E. Household Accessibility to Heat Refuges: Residential Air Conditioning, Public Cooled Space, and Walkability. *Environ. Plan. B Urban Anal. City Sci.* **2017**, 44 (6), 1036–1055. <https://doi.org/10.1177/0265813516657342>.
- (309) U.S. Energy Information Administration - EIA - Independent Statistics and Analysis. <https://www.eia.gov/consumption/residential/data/2020/index.php?view=microdata> (accessed 2024-11-06).
- (310) Bureau, U. C. *American Housing Survey (AHS)*. Census.gov. <https://www.census.gov/AHS> (accessed 2023-08-25).
- (311) 2020 RECS Technical Documentation. **2020**.
- (312) SLOPE. <https://maps.nrel.gov/slope> (accessed 2024-10-24).
- (313) Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; Kegelmeyer, W. P. SMOTE: Synthetic Minority Over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, 16, 321–357. <https://doi.org/10.1613/jair.953>.
- (314) He, H.; Garcia, E. A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, 21 (9), 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.
- (315) Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Handling Imbalanced Datasets: A Review. *GESTS Int. Trans. Comput. Sci. Eng.* **2005**, 30, 25–36.
- (316) Fernández, A.; García, S.; Galar, M.; Prati, R. C.; Krawczyk, B.; Herrera, F. *Learning from Imbalanced Data Sets*; Springer International Publishing: Cham, 2018. <https://doi.org/10.1007/978-3-319-98074-4>.
- (317) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, 29 (5), 1189–1232.
- (318) Breiman, L.; Friedman, J.; Olshen, R. A.; Stone, C. J. *Classification and Regression Trees*; Chapman and Hall/CRC: New York, 2017. <https://doi.org/10.1201/9781315139470>.
- (319) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer: New York, NY, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
- (320) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; KDD '16; Association for Computing Machinery: New York, NY, USA, 2016; pp 785–794. <https://doi.org/10.1145/2939672.2939785>.
- (321) Power Outage | Where. <https://community.fema.gov/ProtectiveActions/s/article/Power-Outage-Where> (accessed 2024-11-06).
- (322) Krieger, J.; Higgins, D. L. Housing and Health: Time Again for Public Health Action. *Am. J. Public Health* **2002**, 92 (5), 758–768. <https://doi.org/10.2105/AJPH.92.5.758>.
- (323) Swope, C. B.; Hernández, D. Housing as a Determinant of Health Equity: A Conceptual Model. *Soc. Sci. Med.* **2019**, 243, 112571. <https://doi.org/10.1016/j.socscimed.2019.112571>.
- (324) Hernández, D.; Swope, C. B. Housing as a Platform for Health and Equity: Evidence and Future Directions. *Am. J. Public Health* **2019**, 109 (10), 1363–1366. <https://doi.org/10.2105/AJPH.2019.305210>.
- (325) *Measuring A/C Access To Prepare Against Extreme Heat*. Federation of American Scientists. <https://fas.org/publication/air-conditioning-data/> (accessed 2024-11-06).

- (326) *Weatherization Assistance Program*. Energy.gov.
<https://www.energy.gov/scep/wap/weatherization-assistance-program> (accessed 2024-11-06).
- (327) *Low Income Home Energy Assistance Program (LIHEAP)*.
<https://www.acf.hhs.gov/ocs/programs/liheap> (accessed 2024-11-06).
- (328) States Call for Congress to Restore Funding for LIHEAP: ABout 1.4 Million Households Could Be Cut from the Program. <https://neada.org/wp-content/uploads/2024/01/pr-recordhbehind.pdf> (accessed 2024-11-06).
- (329) *LIHEAP Fact Sheet*. <https://www.acf.hhs.gov/ocs/fact-sheet/liheap-fact-sheet> (accessed 2024-11-06).
- (330) White, R. H.; Anderson, S.; Booth, J. F.; Braich, G.; Draeger, C.; Fei, C.; Harley, C. D. G.; Henderson, S. B.; Jakob, M.; Lau, C.-A.; Mareshet Admasu, L.; Narinesingh, V.; Rodell, C.; Roocroft, E.; Weinberger, K. R.; West, G. The Unprecedented Pacific Northwest Heatwave of June 2021. *Nat. Commun.* **2023**, 14 (1), 727. <https://doi.org/10.1038/s41467-023-36289-3>.
- (331) Silberner, J. Heat Wave Causes Hundreds of Deaths and Hospitalisations in Pacific North West. *BMJ Br. Med. J. Online* **2021**, 374. <https://doi.org/10.1136/bmj.n1696>.
- (332) *Results of heat mapping project show inequitable impact of hotter summers, will inform actions by King County and City of Seattle - King County, Washington*.
<https://www.kingcounty.gov/en/legacy/elected/executive/constantine/news/release/2021/june/23-heat-mapping-results> (accessed 2024-11-06).
- (333) *Impacts of the 2021 heat dome on emergency department visits, hospitalizations, and health system operations in three hospitals in Seattle, Washington - PMC*.
<https://pmc.ncbi.nlm.nih.gov/articles/PMC10799240/#emp213098-bib-0025> (accessed 2024-11-06).
- (334) *AG Ferguson, Sen. Nguyen propose legislation to protect Washingtonians from utility shutoffs during extreme heat | Washington State*. <https://www.atg.wa.gov/news/news-releases/ag-ferguson-sen-nguyen-propose-legislation-protect-washingtonians-utility> (accessed 2024-11-06).
- (335) *CHAPTER 4 RE RESIDENTIAL ENERGY EFFICIENCY - 2021 WASHINGTON STATE ENERGY CODE - RESIDENTIAL PROVISIONS*.
<https://codes.iccsafe.org/content/WAERC2021P1/chapter-4-re-residential-energy-efficiency> (accessed 2024-11-06).
- (336) *Reports to Congress | LIHEAP Performance Management*.
https://liheappm.acf.hhs.gov/reports_to_congress (accessed 2024-11-06).
- (337) *DP04: Selected Housing Characteristics - Census Bureau Table*.
<https://data.census.gov/table/ACSDP5Y2022.DP04?g=040XX00US36&tid=ACSDP5Y2022.DP04> (accessed 2024-11-06).
- (338) Table HC7.3 Air Conditioning in U.S. Homes, by Year of Construction, 2020.
<https://www.eia.gov/consumption/residential/data/2020/hc/pdf/HC%207.3.pdf> (accessed 2024-11-06).
- (339) *2024 heat mortality annual report in NYC | Environment and Health Data Portal*. Environment & Health Data Portal. <https://a816-dohbep.nyc.gov/IndicatorPublic/data-features/heat-report/> (accessed 2024-11-06).
- (340) Bauman, A. *NYC landlords would be required to provide air conditioning for tenants under newly introduced bill - CBS New York*. <https://www.cbsnews.com/newyork/news/nyc-air-conditioning-bill/> (accessed 2024-11-06).
- (341) *Cooling Assistance Benefit – ACCESS NYC*. <https://access.nyc.gov/programs/cooling-assistance-benefit/> (accessed 2024-11-06).

- (342) Moya, E. M.; Chávez-Baray, S. M.; Wood, W. W.; Martinez, O. Nuestra Casa: An Advocacy Initiative to Reduce Inequalities and Tuberculosis along the US-Mexico Border. *Int. Public Health J.* **2016**, *8* (2), 107.
- (343) Busby, J. W.; Baker, K.; Bazilian, M. D.; Gilbert, A. Q.; Grubert, E.; Rai, V.; Rhodes, J. D.; Shidore, S.; Smith, C. A.; Webber, M. E. Cascading Risks: Understanding the 2021 Winter Blackout in Texas. *Energy Res. Soc. Sci.* **2021**, *77*, 102106.
<https://doi.org/10.1016/j.erss.2021.102106>.
- (344) Kemabonta, T. Grid Resilience Analysis and Planning of Electric Power Systems: The Case of the 2021 Texas Electricity Crises Caused by Winter Storm Uri (#TexasFreeze). *Electr. J.* **2021**, *34* (10), 107044. <https://doi.org/10.1016/j.tej.2021.107044>.
- (345) Flores, N. M.; McBrien, H.; Do, V.; Kiang, M. V.; Schlegelmilch, J.; Casey, J. A. The 2021 Texas Power Crisis: Distribution, Duration, and Disparities. *J. Expo. Sci. Environ. Epidemiol.* **2023**, *33* (1), 21–31. <https://doi.org/10.1038/s41370-022-00462-5>.
- (346) *Compound Climate and Infrastructure Events: How Electrical Grid Failure Alters Heat Wave Risk | Environmental Science & Technology*. <https://pubs.acs.org/doi/10.1021/acs.est.1c00024> (accessed 2024-11-06).
- (347) Li, X. ‘Cathy’; Zhao, L.; Qin, Y.; Oleson, K.; Zhang, Y. Elevated Urban Energy Risks Due to Climate-Driven Biophysical Feedbacks. *Nat. Clim. Change* **2024**, *14* (10), 1056–1063.
<https://doi.org/10.1038/s41558-024-02108-w>.
- (348) Buonocore, J. J.; Salimifard, P.; Magavi, Z.; Allen, J. G. Inefficient Building Electrification Will Require Massive Buildout of Renewable Energy and Seasonal Energy Storage. *Sci. Rep.* **2022**, *12* (1), 11931. <https://doi.org/10.1038/s41598-022-15628-2>.