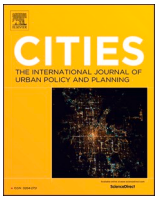
Cities 153 (2024) 105170

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities



The association between urban density and multiple health risks based on interpretable machine learning: A study of American urban communities

Zerun Liu [a], Chao Liu [b,*]

[a] Tandon School of Engineering, New York University, 6 MetroTech Center, Brooklyn, NY 11201, United States of America
[b] Department of Urban Planning, College of Architecture and Urban Planning, Tongji University, Shanghai Tongji Urban Planning and Design Institute Co. Ltd., Room 203, Wenyuan Bulding, No. 1239, Siping Road, Shanghai 200092, China

ARTICLE INFO

Keywords:
Urban density
Interpretable machine learning
Public health
Built environment
Chronic disease
Population density

ABSTRACT

With the growing complexity in urban areas, cities have become unprecedentedly intricate systems. This paper aims to develop interpretable machine learning (ML) approaches to unravel the sophisticated associations. In a case study of American urban communities, we apply interpretable ML methods to identify the associations between urban density and multiple health risks. We define urban density from three dimensions of population, built environment, and activity and measure multiple health risks based on categories of physical diseases, mental diseases and health burden. Initially, we conduct cluster analysis to control socioeconomic variables and select study samples. Then we build several ML models of multiple linear regression, decision trees, random forests, and extreme gradient boosting. Interpretable methods, including feature importance, partial dependence plots, individual conditional expectations, and Shapley additive explanations, are used to interpret the models by identifying important factors, non-linear relationships, the interactions between variables. The results show the advantages of interpretable ML methods in efficiency and transparency. Our findings verify that the associations between urban density and multiple health risks are complicated. The similarities and differences between various health risks provide valuable evidence on key factors, thresholds, and pathologic characteristics that can guide the healthy and sustainable development of cities.

1. Introduction

Urban density has recently garnered significant attention, primarily due to the increasing global urban populations. Generally, urban density refers to the concentration of people, buildings, and activities within a given area of a city. Density, embodying both physical and cultural aspects of urban space and bearing significance in economics and society, serves as an effective tool for quantifying and measuring cities (Bettencourt et al., 2007). However, one of the most controversial debates about urban density is that while high density contributes to great efficiency, diversity, and sharing of resources, compactness within cities diminishes the quality of an individual's life, given the increase of disease, pollution, and crime (Glaeser et al., 2001; Pan et al., 2013). Recent research, particularly during the pandemic, has exacerbated this contradiction. Evidence shows that though dense urban areas have better medical conditions and infrastructure systems, their high activity intensity, frequent human mobility, and close social contact cause the rapid explosion of cases. Still, no unified and convincing research shows the mechanisms and rules of urban density.

In fact, the debate over the impact of urban density has been discussed since the inception of cities. As a product of urbanization, the concept of spatial density originates from agglomeration attributes since the birth of cities. Initially, dense urban centers were pivotal for innovation and wealth concentration, facilitating the transition from rural to urban societies (Becker et al., 1999; Bettencourt & West, 2010; Bettencourt et al., 2007; Milgram, 1974). However, the Industrial Revolution brought rapid population increases and dense, often unhealthy living conditions, highlighting the need for urban development control. This led to the 1848 Public Health Act, which marked the beginning of government-managed urban planning (Townshend & Lake, 2009). The 19th and 20th centuries saw the rise of modern urban planning and movements like the Garden City, proposed by Ebenezer Howard, which aimed to balance urban amenities with green spaces, addressing urban sprawl (Howard, 1902). This movement emphasizes the importance of controlling urban sprawl and managing urban density. Along with the 21st century's coming, deepening urbanization and globalization bring

* Corresponding author.
E-mail addresses: zl3280@nyu.edu (Z. Liu), liuchao1020@gmail.com (C. Liu).

https://doi.org/10.1016/j.cities.2024.105170
Received 10 April 2023; Received in revised form 30 May 2024; Accepted 1 June 2024
Available online 25 July 2024

new challenges of accommodating this urban growth while ensuring livability, efficiency, and sustainability. Still, there are few clear and confirming conclusions indicating the roles of urban density in modern cities and how to appropriately guide the development of dense urban areas.

In a range of fields such as physics, economics, sociology, geography, architecture, and urban planning, the concept of density holds significant importance. Its definition varies, and this variability is particularly pronounced in urban studies, where its implications can differ significantly. Broadly, density encompasses the integration of social interaction, economic activities, energy flow, ecological environment, culture, history, and more. In a narrow sense, especially from the perspective of practice in urban planning and construction, urban density usually refers to population density and building density, which are commonly discussed together with land use, accessibility, road network, etc. (Carlino et al., 2007; Dovey & Pafka, 2014). Recent studies shed light on how social tie density and population density influence the efficient generation of ideas and augment productivity in cities (Pan et al., 2013). And more scholars pay attention to the interactions within networks and the spatial agglomeration of human mobility (González et al., 2008; Levinson, 2012; Louf & Barthelemy, 2014; Simini et al., 2012). There is an urgent need for a general model to explain how density covering population, built environment, and mobility generatively relate to urban development, ranging from wealth and innovation to crime and disease. Thus, in this study, we measured urban density in three aspects: population density (the distribution intensity of population in unit space like area and room), built environment density (the development intensity of urban space), and activity density (the spatial distribution intensity of activities and facilities). By synthetically considering these aspects, we can comprehensively measure and understand the cities.

Among the extensive research on urban density, studies focusing on health risks are particularly noteworthy. Health risks, defined as the probability of adverse health outcomes, encompass a broad spectrum of elements potentially harmful to an individual's physical, mental, or emotional well-being (WHO, 2009). Generally, health risks can be divided into three categories: (1) Physical diseases, which include chronic diseases such as heart diseases, cancers, and diabetes, and infectious diseases such as flu and AIDS. (2) Mental diseases, which include depression and schizophrenia. (3) Health burden, which refers to significant public health hazards that increase the probability of multiple diseases such as obesity, sleep deprivation, anxiety, smoking, and alcohol use. Previous literature has proven that health risks can arise from various sources, including genetic predispositions, environmental factors, socioeconomic conditions, lifestyles, and behaviors. Clearly, urban density is closely related to all these aspects, necessitating an investigation and explanation of its association with multiple health risks.

However, there are different voices about the relationships between urban density and health risks. Some have concluded that dense urban areas might cause health problems because of the limited resources and overcrowded space. Others have concluded that density within cities could bring more life convenience and abundant facilities that improve health and medical care. Moreover, a number of researchers believe the relationships are complex or non-linear. Still, there is no unified clear conclusion describe the association between urban density and health risks. Therefore, in this study, we aim to identify the complex relationships between urban density and multiple health risks. Utilizing interpretable machine learning, we measure multi-dimensional urban density and compare its associations with various health risks, offering comprehensive insights into the relationship between urban density and health risks. The interpretable machine learning methods employed are valuable for detecting complex and non-linear relationships, applicable to other research endeavors aimed at uncovering non-linear relationships.

This study contributes to the previous theories and literature from the following aspects: The results would broaden the comparison between various health risks and deepen the understanding of urban density. The methods would promote the development of interpretable ML and its applications. The primary novelty of method lies in the application of these models to uniquely dissect and illuminate the health implications of different urban density measures. This application yields critical insights that are instrumental in advancing urban health research and practice. Generally, our research could facilitate progress in both methodology and knowledge of health problems in urban areas and support the control practice of urban density, which benefits the healthy, livable, and sustainable development of cities.

The research flow is depicted in Fig. 1. In a case study based in the United States, we explore various health risks at the census tract level. With the help of the K-means clustering algorithm, study samples are selected by grouping to control socioeconomic factors. To identify the non-linear correlations between the density indicators and multiple health risks, several machine learning models, including multiple linear regression (MLR), decision tree (DT), random forest (RF), and extreme gradient boosted trees (XGBoost), are built for the selected samples. We then conduct interpretable methods at global and local scales to reveal the key influencing density measurements, describe the complicated relations, and understand the interactions between different factors. The results of this case study provide preliminary conclusions about the comparison of different health risks, potential mechanism of urban density, and applicable methods of interpretable machine learning.

This paper is organized as follows: Section 2 provides a review of the literature on urban density and its association with various health risks. Section 3 describes the methodology and data utilized in this study, including details on data processing. Section 4 presents the results of the case study, where, with the aid of interpretable machine learning analysis, we elucidate key driving factors, describe non-linear relationships, identify critical thresholds, and reveal the interactions among various features. In Section 5, we discuss the primary findings from this analysis, exploring potential underlying reasons. Finally, Section 6 concludes the paper by summarizing our hypotheses and the contributions derived from this research.

2. Literature review

In order to effectively compare existing literature, we select several pivotal studies based on the principles of diversity and comprehensiveness, covering various types of health risks, regions, and methodologies. Table 1 is constructed to summarize these studies, highlighting aspects such as study regions, health outcomes, density factors, methodologies, primary findings, and the correlations between urban density and health risks. Although numerous cutting-edge discoveries have been made regarding different health risks, there remains a lack of universal, consistent conclusions. The outcomes varied, showing positive, negative, non-linear, super-linear, and multifaceted relationships. This divergence may be attributed to differences in health risk types, geographic regions, assumed relationships, and methodological approaches.

Firstly, multiple types of health risks produce mixed results in different countries and regions. As for chronic diseases, many scholars have uncovered the significant influence of density factors, particularly focusing on population density. Similar conclusions are drawn for typical chronic diseases, including heart diseases, high blood pressure, and diabetes, that the increase in population density would lead to a decrease in the incidence rate (Griffin et al., 2013; Konishi et al., 2020). While some studies get the inverse conclusions in dense urban areas (Li et al., 2022; Yang & Hsieh, 1998), indicating that environmental factors had cumulative effects on those diseases. As for infectious diseases, population density and POI density play the leading roles in morbidity and mortality. The outbreaks of infectious diseases usually start earlier in densely populated regions. But the conclusions are more complex, with unignorable disparities between different regions and cases (Yip et al., 2021; Li, Peng, et al., 2021; Hu et al., 2021; Zhang, 2020; Mollalo

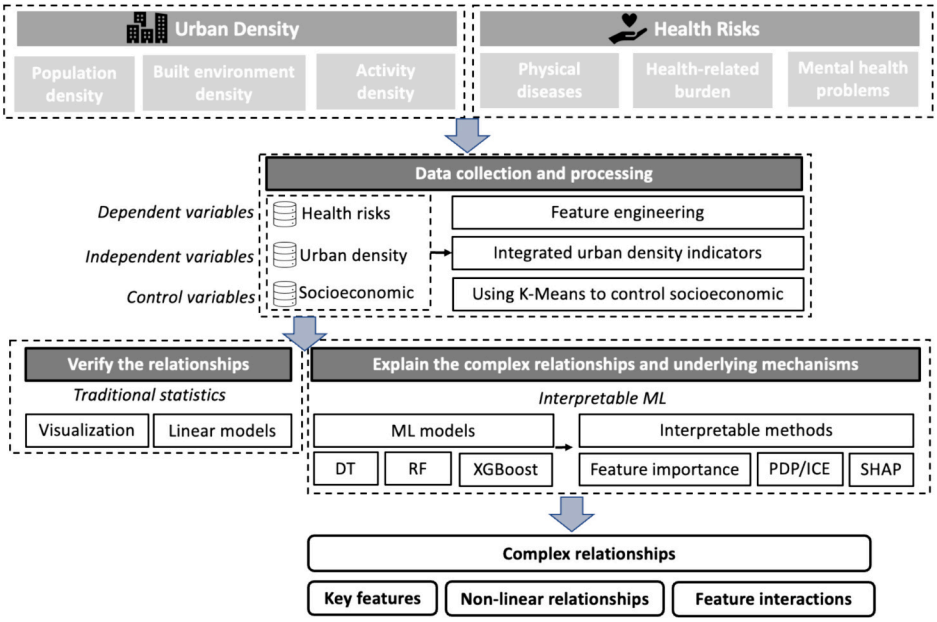


Fig. 1. Research flow.

et al., 2020). As for mental diseases, open space and population density are getting more attention, which are proven to be important factors affecting the feelings of people toward environments (Gruebner et al., 2017; Melis et al., 2015). As for typical health burdens such as obesity, anxiety, and sleep reduction, the majority of studies indicate that population density, open space, and crowd within communities significantly influenced the behaviors of people living in urban areas, causing composite cumulative health outcomes (Ewing et al., 2003; Lopez, 2007; Rundle et al., 2007). Apparently, the majority of studies only focus on some specific health risks. Few have conducted a comprehensive study on various health risks. The findings of multiple health risks drawn from different regions are chaotic, lacking comprehensive comparisons and clear clues. Therefore, it is necessary to conduct a comprehensive comparative study of various types of health risks in the same region.

Although existing research posits that the relationships between urban density factors and health risks are multifaceted and vary by region, three predominant hypotheses have emerged to describe the associations between urban density and health outcomes: (1) A positive linear relationship suggests that increasing urban density correlates with higher health risks, while lower density is associated with fewer health issues. (2) A negative linear relationship implies that higher urban density can decrease health risks, whereas lower density may increase them. (3) A non-linear relationship indicates that both very low and very high urban densities can negatively affect health. Historically, research predominantly focused on linear relationships, overlooking the complexities of urban density effects on health. Recent studies, particularly in China, exploring obesity and walkability (Lu et al., 2017; Yin et al., 2022), advocate for a non-linear perspective. Yet, there remains a gap in understanding these relationships in depth, particularly in explaining the mechanisms and interactions at play. This study aims to bridge this gap, enhancing our understanding of the nuanced relationships between urban density and health risks.

One of the biggest challenges that prevent previous research from thoroughly exploring the mechanism of model and complex correlations is the methodology. The large volume of data, the diverse types of data structures, and the complex relationships between factors are beyond the capacity of traditional methods. General statistical methods like correlation analysis, and spatial analysis cannot effectively deal with data from multiple sources and uncover the non-linear relationships. The emergence of machine learning enables the processing of big and diverse data and the detection of unexpected rules in complicated structures. However, the difficulty in interpreting the decision-making process in such black-box models becomes one of the main complaints about machine learning methods. To overcome this weakness and improve the transparency of black-box models, interpretable methods are introduced in related applications.

Interpretable machine learning methods, also known as explanatory algorithms, aim to transform complex computational models into formats comprehensible to humans (Bi et al., 2020; Molnar, 2022). According to the different stages of the occurrence of explanation, interpretability can be divided into interpretable models and post-interpretable methods for uninterpretable models (Molnar, 2022). Common interpretable models, also called white-box models, mainly include linear regression, logistic regression, RuleFit, Naïve Bayes, and k-nearest neighbors. Due to the relatively simple algorithm behind these models, the decision-making process can be easily recognized. However, other black-box models like ensemble learning and neural networks are too complex to describe in simple language. Thus, extra methods are needed to demonstrate special layers in the intricate models. Two types of interpretable methods are designed for those black models: the global model-agnostic methods and the local model-agnostic methods (Molnar, 2022). The global interpretability looks at all the possible inputs and outputs and the relationships between them, literally explaining the global views. Local interpretability focuses on the comprehension of special samples, that is, more individual, unique, and local views. Permuted feature importance, partial dependence plot (PDP), individual conditional expectation (ICE), and global surrogate are widely used for the global explanation. As for local interpretability, local interpretable model-agnostic explanations (LIME) and Shapley additive explanations (SHAP) are the most popular in general research. The aforementioned interpretable methods have been employed in various fields, including transportation, medicine, clinical diagnosis, disease prediction, and environmental detection, often in conjunction with ensemble tree algorithms. Considering the weaknesses and strengths of different explainable methods, the integrated use of multiple methods might be the most suitable choice for a comprehensive study.

Overall, most existing studies only focus on specific health problems and have inconsistent conclusions about the associations between urban density and health risks. There are research gaps in the comprehensive research on multiple health risks and explanations about complex

Table 1
Summary of methods in related research.

Source	Study area	Health risks	Main factors (density related)	Methods	Main findings	Associations between density and health risks
Chronic diseases						
(Hipp & Chalise, 2015)	U.S.	Diabetes	**Population density**, percentage population that cycles or walks to work	Geographically weighted regression	Among the environment-related variables examined, the percentage of the population commuting to work by cycling or walking was the sole significant factor associated with diabetes, and the strength of this correlation varied across different regions of the country.	Complex
(Li et al., 2022)	Wuhan, China	Ischemic heart disease (IHD) risk	**Population density**, floor-area ration, NDVI, facility	Advanced Bayes Bayesian CAR model	Communities with a higher open space ratio, increased greenery, improved access to healthy food, and the presence of sports facilities were associated with a reduced risk of IHD.	Positive
(Bassolas et al., 2019)	U.S.	Sudden ischemic death	**Activity density**, structure of mobility	Pearson and Spearman correlation, LOESS	Cities with higher mobility levels were more likely to widely use public transport, provide better walking environment, cause lower per capita pollutant emissions, and lead to better health evaluation.	Negative
(Xie & Zhu, 2020)	China	Stroke risk	**Net population density, employment density**, architectural characteristics, facilities, open space, transport, land-use pattern	Bayesian CAR model	The research integrates stroke-related built environment factors into a framework, demonstrating that higher net population and building densities are positively linked to stroke risk. The findings suggest that compact urban developments may not be universally beneficial for health in transitional cities, underscoring the need for a critical reassessment of their suitability.	Positive
Infectious diseases						
(Hamidi et al., 2020)	U.S.	Covid-19	**Activity density**, metropolitan population	Multilevel linear model	The spread of COVID-19 is more influenced by connectivity than by density, with larger cities being more prone to outbreaks but densely populated areas showing lower mortality rates. Air travel frequency has a negligible impact on infection rates.	Complex
(Li, Ma, & Zhang, 2021)	China	Covid-19	Centrality, **POI density**, **population density**	Mixed geographically weighted regression	The results showed that 28.63 % of cities had low infection rate which was related to spatial agglomeration. The POI density around the railway station, the distance from public transport, and the number of flights from Hubei Province all affect the spread of COVID-19.	Mostly positive
(Pan et al., 2013)	North America and Europe	AIDS/HIV	**Population density**, social tie intensity	Data visualization, SI model	There was super-linear relationship between social tie density and infectious diseases.	Super-linear
(Rader et al., 2020)	China and Italy	COVID-19	**Population density**, mobility	Data visualization, correlation analysis	In crowded cities, epidemics spread more widely over time. And crowded cities had a higher overall incidence rate than cities with smaller population density.	Positive
Health burden						
(Yin et al., 2022)	China	Overweight (BMI)	**Residential density**, physical activity time	Mixed-effects logistic regression model	This study found a positive association between residential density and the prevalence of overweight among urban Chinese adolescents. These findings highlight the need for additional research to explore the characteristics of urban environments that relate to obesity in Chinese adolescents and to understand the underlying mechanisms of this relationship	Positive
(Rundle et al., 2007)	New York, U.S.	Obesity (BMI)	Land use, the density of metro and bus stations, **population density**	Multilevel analysis	Land use, the density of metro and bus stations, and population density were negatively correlated with BMI.	Negative
(Xu et al., 2010)	China	Obesity (waist-hip ratio)	**Regional built area, regional and local population density**, land use diversity, bus stops, distance to business center	Gradient boost decision trees	Neighborhood population density shows a U-shaped correlation with waist-hip ratio, with a threshold effect observed at the regional level. Built environment factors, notably neighborhood density and regional built area, are more significant in predicting waist-hip ratio than socio-demographic	U-shape

(*continued on next page*)

Table 1 (*continued*)

Source	Study area	Health risks	Main factors (density related)	Methods	Main findings	Associations between density and health risks
					factors, among which household income is the most influential.	
Mental health						
(Melis et al., 2015)	Italy	Metal health	**Density**, public transport accessibility, facility convenience, green space and open space accessibility	Multiple linear regression	The accessibility of public transport and urban density were only related to mental health, which played relatively weak protective roles. For people who spend more time in the community, the built environment had a strong impact on mental health.	Weakly negative
(Chan et al., 2021)	Hong Kong, China	Anxiety	**Living density**, house characteristics	Logistic regression	Living in medium-density areas is associated with a lower anxiety risk compared to high-density areas. This risk decreases further in low-density areas. Additionally, residing in low-density areas significantly reduces stress risk.	Complex
(Ruiz-Tagle & Urria, 2022)	Chile	Mental health (depressive symptoms)	**Household density**, housing conditions	Descriptive statistics	An increase in overcrowding correlates with more depressive symptoms, but decreasing overcrowding doesn't significantly affect mental health, particularly when the change involves the number of bedrooms.	Asymmetric

relationships. Leveraging advanced interpretable ML methods and multi-sourced data, this study is intended to fill research gaps. It utilizes interpretable ML techniques to unravel the intricate associations between urban density and various health risks, focusing on case studies in American metropolitan areas. The results can help urban planners and decision-makers reveal more details about the mechanisms of urban density to promote sustainable and healthy development, Eventually, this study can extend the theories and empirical research of urban density and public health.

3. Method and data

3.1. Data and variables

We gather the data at the level of census tracts in the contiguous United States, in which Alaska and Hawaii are excluded because their unique environmental conditions and population dynamics do not align with those of the continental states. Designed for the purpose of taking a census, the census tract usually has a relatively homogeneous population size, characteristics, and living environment, which is beneficial to conducting density research. Besides, as the smallest territorial entity, a census tract can provide huge amounts of data on population, environment, economics, and public health. In total, data from over 700,000 census tracts are gathered for this research.

In this study, multiple data sets of health risk, urban density, and socioeconomic data are collected from government, open platforms, and social organizations. According to the different categories of health risks, we focus on the seven most influential cases of coronary heart disease (CHD), cancer (except skin), diabetes, current asthma, COVID-19, obesity, sleep reduction, and mental health problems. The COVID-19 data are published by Johns Hopkins University (JHU) and gathered in a COVID-19 data project (Dong et al., 2020). Other health data are collected from the PLACE project in 2018, funded by the Centers for Disease Control and Prevention (CDC) and the Robert Wood Johnson Foundation (RWJF) (Centers for Disease Control and Prevention, 2016-2022). To cover the three dimensions of population density, built environment density, and activity density, we gather population data from the 2010 Census combined with the 2014–2018 America Community Survey (ACS), summarize built environment data from the National Land Cover Database (NLCD) (Philippa Clarke, 2001–2016), and calculate activity density on the foundation of POI data from the Open Street Map (OSM) (Foundation: Cambridge, 2020). And we calculate the integrated urban density indicators by synthesizing multidimensional density factors with the help of the entropy method. In addition, we monitor socioeconomic data from the perspectives of age, poverty, income, and minority, which is from the database of the Agency for Toxic Substances and Disease Registry (ATSDR) Social Vulnerability Index (SVI) provided by the CDC (Geospatial Research, 2021). All the data and corresponding variables are listed in Table 2.

Table 2
Summary of variables and data sources.

Categories	Variables	Data sources
Health risks		
Diseases		
Chronic	Coronary heart disease (CHD), cancer, diabetes, current asthma	CDC
Mental health	Mental health	CDC
Health burden		
Obesity	Obesity	CDC
Lack of sleep	Sleep<7	CDC
Urban density		
Built environment density		
Open space	Open space	NLCD Land Cover
Intensity	Low-medium-high-intensity	NLCD Land Cover
Population density		
Population density	Population density (PD)	Census/ASC
The degree of crowd	EP_CROWD	Census/ASC
Activity density		
POI density	POI density (d_poi)	OSM
Socioeconomic factors		
Age	Age > 65 (AGE65), age < 17 (AGE17)	Census/ASC/ ATSDR SVI
Race	The ratio of minority (EP_MINRTY)	Census/ASC/ ATSDR SVI
Income	The ratio of poverty (EP_POV), Per capita income (EP_PCI)	CDC/ATSDR SVI
Employment	The ratio of unemployment (EP_UNEMP)	CDC/ATSDR SVI

3.2. Study area selection

To control for socioeconomic variables and focus on the impact of density factors, we employ grouping methods for sample selection. Cluster analysis, an unsupervised pattern recognition technique, effectively partitions targets into homogeneous clusters, identifying significant similarities. One of the most common and efficient cluster algorithms is K-means clustering. Thus, we apply K-Means analysis for optimal socioeconomic clustering, using the elbow method to determine the ideal number of clusters. Although clustering algorithms are sensitive to data and somewhat subjective, we scrutinize feature distributions to validate the appropriateness of diverse clustering results. More details are provided in Appendix 1. The results of contiguous United States without Alaska and Hawaii are shown in Fig. 2. We ultimately select 14,304 census tracts labeled as 'cluster_2' for our study samples, predominantly situated in urban areas characterized by high urbanization, active economic activities, developed living environments, and relatively uniform income and consumption levels.

3.3. Model building

This paper explores various methods to elucidate the correlations between the incidence rates of multiple health risks (dependent variables) and urban density factors (independent variables). We first use the forward stepwise multivariable linear regression model (MLR) to build models of important variables. Recognizing the absence of a perfect linear relationship in our data, we subsequently developed more intricate machine learning (ML) models. As a non-parametric tree structure model, the decision tree (DT) could closely explain decision rules and consequences step by step. The classification and regression tree (CART) is more feasible and applicable for the regression task, which is applied in this research. Still, DT had limitations in stability, generalizing, efficiency, and predicting complex continuous values. Thus, we introduce integration algorithms for random forest (RF) models and extreme gradient boosting (XGBoost) models to reduce the risk of overfitting, lower the sensitivity to extreme data, improve generalization and efficiency, and enhance the robustness of our models (Fig. 3).

3.4. Interpretable ML

Although the accuracy and capability of ML models have greatly improved, we knew little about the detailed decision-making process of their internal workings. Fortunately, many interpretable methods are developed to explain how the consequences are produced. And this paper pays attention to the methods of feature importance, partial dependence plot (PDP), individual conditional expectation (ICE), and Shapley values. The widely used measurements for feature importance are model-based feature importance, permutation feature importance, and SHAP value feature importance. Among them, the model-based feature importance depends greatly on the algorithm used by the model, making the comparison between different models difficult. Therefore, permutation feature importance and SHAP value feature importance are chosen in this research. By randomly shuffling a single feature, permutation feature importance defines the importance of each feature as the decrease in the model score. Shapley values are from cooperative game theory, and Shapley additive explanations (SHAP) are based on the classic Shapley values, which interpreted the model output by calculating the mean marginal contribution of each feature across all the features. As for local samples, Shapley values could demonstrate the degree and direction of feature influence. As for global models, summing the average of the absolute Shapley values of features could provide supporting evidence for the features' contributions to the outcome. PDP shows the mean marginal effect of certain features by marginalizing other values of estimators, which could reflect the linear or non-linear relationships between inputs and outputs. ICE displays the dependence between the targets and features of one instance, and it could be seen as the decomposition of PDP. Meanwhile, SHAP dependence plots, serving as an alternative to PDP, offer more detailed insights by considering feature interactions, derived from subtracting the main individual effect.

4. Result

4.1. Statistical results

For further explorations of complex relationships and to identify the influence of multi-dimensional density factors, we build several models of MLR, decision tree, random forest, and XGBoost to study the interactions between dependent variables of incidence rates and independent variables of multi-dimensional density factors. Using the methods of grid search and cross-validation, we determine the best parameters for each model and conducted them on our training dataset. We compare the performance of different models through the indicators

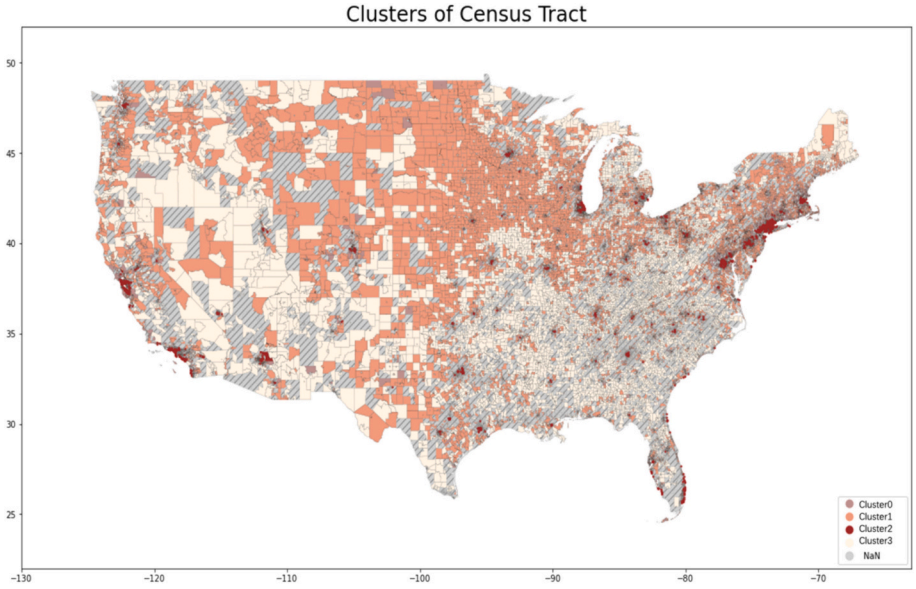


Fig. 2. Clustering results (label "Cluster2" was selected as study areas).

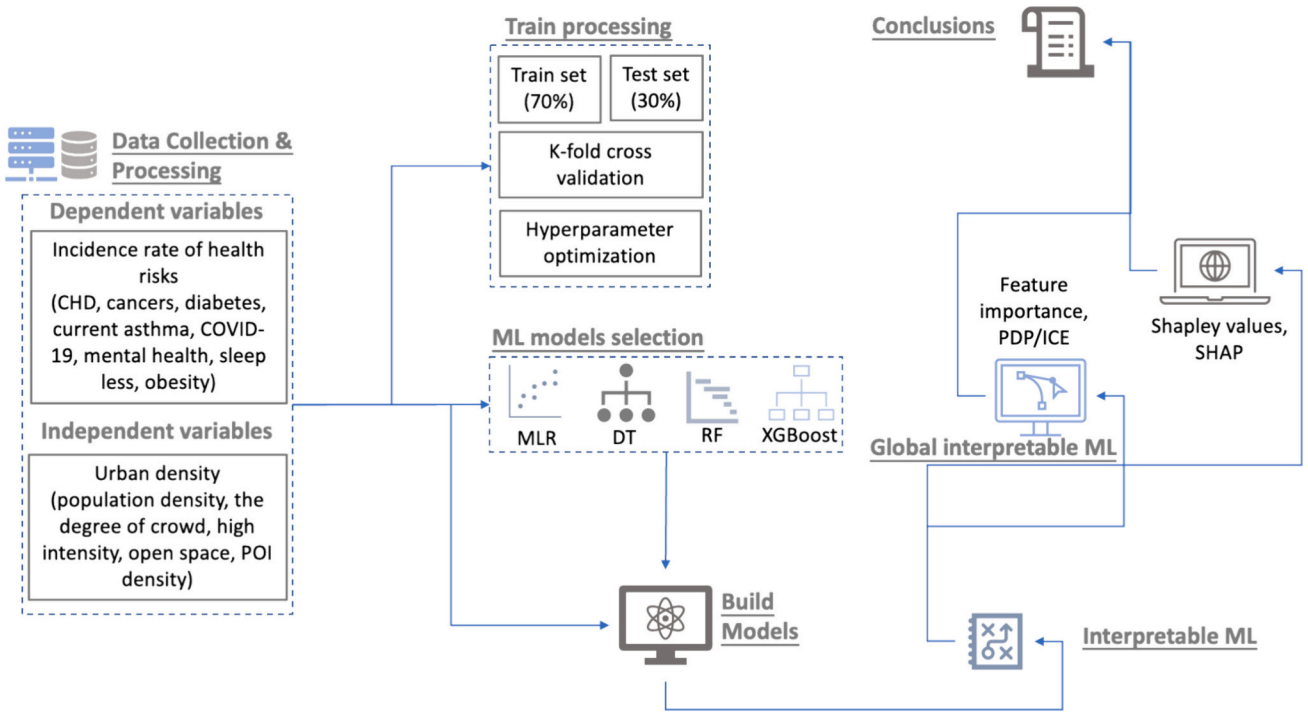


Fig. 3. Model building framework.

of MAE, MSE, RMSE, and R-Square. Table 3 shows the results of our trained models. It can be seen that from the simple MLR and decision tree models to the more complicated random forest and XGBoost models, the prediction accuracy continue to improve. And the errors in chronic diseases are much smaller than those in infectious diseases and mental health problem. However, the overly complex models are usually prone to overfitting, worsening the ability to generalize. In our research, the random forest models achieve the optimal balance of accuracy and generalization in fitting different health risks. Therefore, we choose the trained random forest models as the foundation for subsequent explanatory analysis.

4.2. ML interpretation

To understand the decision-making process of ML models, we employ global and local explanatory methods to interpret our best models. Specifically, permutation feature importance, PDP, ICE, and SHAP analysis are used in this research to help us reveal the key factors, uncover the non-linear relationships, and understand the interactions between variables (Table 4).

4.2.1. Key factors

Based on the results of the trained random forest models, we measure the importance of multi-dimensional factors with the help of permutation feature importance and Shapley values. Permutation feature importance can provide insights into the global contributions of each feature by detecting the disturbance to the performance of models in the process of replacing each feature (Altmann et al., 2010). While the summary of Shapley values can introduce the local views from each sample, in which the average of absolute Shapley values reflect the contributions of the features to the outcome variables, each point could present the specific contribution of each feature for a certain sample. The results show that the ranking of important features output by different methods was almost the same, making it more convincing. Homogenous health risks manifest similarities. As for the chronic diseases of CHD, diabetes, and cancer, population density and high intensity are the primary features. The distribution of samples' Shapley values show that the points with higher population density (PD) are more uniformly concentrated near the y-axis, indicating that the positive and negative effects of high population density were not significant. The points with higher intensity are mostly on the left side of the y-axis, indicating that higher intensity had a negative impact on the incidence rates. But POI density plays a greater role in current asthma, and points with higher POI density are more likely to be located on the right side of the y-axis. As for infectious disease, the most important features are population density and POI density, which are similar to the results of obesity. And points with higher feature values are mostly on the negative side. Lack of sleep and mental health have more to do with open space, population density, and crowd degree. The higher feature values of open space and population density remain more on the points on the negative side, which is opposite to the crowd degree (Fig. 4).

4.2.2. Non-linear relationships

Furthermore, we draw the partial dependence plots (PDP) and individual conditional expectations (ICE) of the recognized primary features for each health risk, which helps us intuitively capture the non-linear relationships. ICE reflect the marginal effect of each given feature. PDP can be regarded as the average of ICE for all samples. We observe some interesting results from the plots. As for the chronic diseases of CHD, diabetes, and cancer, as the population density increase, the incidence rate first decrease, slow down, then slightly increase at the end. There are some noteworthy numbers within the threshold of population density. When it is <2000, the downward trend is significant. When it is between 2000 and 4000, the downtrend gradually becomes slight. And the turning point appears around 4000. With the further increase in population density, the incidence rate rise gently, but this growth tendency almost disappears when the population density exceeds 6000. For high intensity, the incidence rate decreases as the ratio of high intensity became higher. And the fit line tends to be linear, with unobvious pits appearing around 0.4. As for current asthma, both population density and POI density show positive relationships with the incidence rate, and performed marginally descending. In contract, the incidence rate of COVID-19 is negatively correlated with population density and POI density, and the declining trend slows down at a high

Table 3
Results of forward stepwise MLR.

Variables	Coefficient	t	p-Value	R2
Obesity				
Constant	1.02E−15	3.64E−14	1	0.457
d_poi	−0.3727	−9.198	<0.001	
log_PD	−0.3882	−9.355	<0.001	
EP_CROWD_x	−0.1729	−5.045	<0.001	
high intensity	0.1399	2.944	0.003	
Sleep < 7				
Constant	6.12E−17	1.82E−15	1	0.227
EP_CROWD_x	0.3331	8.376	<0.001	
high intensity	0.2012	3.621	<0.001	
open space	−0.1178	−2.676	0.008	
log_PD	−0.111	−2.215	0.027	
Current asthma				
Constant	−8.95E−16	−2.69E−14	1	0.235
d_poi	0.3143	7.607	<0.001	
log_PD	0.2271	5.497	<0.001	
CHD				
Constant	−2.70E−16	−8.62E−15	1	0.327
high intensity	−0.2819	−5.436	<0.001	
EP_CROWD_x	−0.1821	−4.906	<0.001	
open space	0.1229	2.991	0.003	
log_PD	−0.106	−2.267	0.024	
Diabetes				
Constant	3.32E−16	1.04E−14	1	0.305
high intensity	−0.3143	−6.341	<0.001	
log_PD	−0.2219	−4.809	<0.001	
EP_CROWD_x	−0.0906	−2.404	0.016	
Cancer				
Constant	−5.48E−16	−1.85E−16	1	0.394
high intensity	−0.3563	−8.451	<0.001	
EP_CROWD_x	−0.2474	−7.037	<0.001	
open space	0.1466	3.87	<0.001	
COVID19				
Constant	−1.334E−16	−1.24E−12	1	0.027
log_PD	0.1811	11.842	<0.001	
high intensity	−0.1269	−7.179	<0.001	
open space	−0.0547	−4.187	<0.001	
Mental health				
Constant	−3.31E−15	−8.97E−14	1	0.062
high intensity	0.1552	3.552	<0.001	
EP_CROWD_x	0.1279	2.928	0.004	

Table 4
Results of decision tree models.

DT	MAE	MSE	RMSE	R2
Obesity	0.61	0.58	0.76	0.40
Sleep < 7	0.78	0.98	0.99	–
Current asthma	0.59	0.68	0.82	0.38
Cancer	0.60	0.62	0.79	0.43
Diabetes	0.59	0.68	0.82	0.38
CHD	0.79	–	–	–
COVID19	0.02	0.001	0.04	–
Mental health	0.71	0.82	0.91	0.06

enough density. More remarkable negative relationships can be observed between the rate of obesity and population density and POI density. As for lack of sleep, the degree of crowding and POI density show positive relationships with the incidence rate, while open space has a negative effect. Similarly, the rate of mental health problems decreases with the increase in open space. But the reducing trend is not clear and powerful enough (Fig. 5).

4.2.3. Interactions of features

However, it is impossible for factors to act independently. More frequently, the interactions between different features also contribute to the model prediction. In order to reveal the details of the decision-making process, we use the Shapley value to calculate the dependence contribution of primary factors to the predicting results. The x-axis represents the true values of selected features, and the y-axis represents the SHAP values that the selected features contribute to the outcomes. When the models are perfect linear relationships, the SHAP dependence plot would show perfect linearity, that is, a certain x value corresponded to a unique y value. However, the actual models often present a diffusion trend, which means that when the x values are consistent or close, the y values are greatly different. The vertical fluctuations are composed of the unpredictable components of the model, data noise, and feature interaction. The first two cannot be effectively captured and difficult to explain. Thus, it is necessary to consider the interactions between features in the prediction results. Based on the results of SHAP interaction analysis, we filter the most active factors and use color to represent the true values of the interactive features, so as to help researchers better observe the vertical fluctuation and reveal the role and interaction effect of the feature (Table 5).

As for the chronic diseases of CHD, diabetes, and cancer, the prediction of incidence rates first decrease and then increase as the population density got higher. And with the increase in the proportion of high intensity, the prediction of incidence rates decrease. The degree of crowding and POI density are the most interactive features of population density. For certain ranges of population density, a higher degree of crowding and POI density correspond to lower incidence rates (Table 6).

As for current asthma, population density and POI density are positively correlated with the incidence rates, and high intensity becomes the most interactive feature for these two primary factors.

The COVID-19 shows the opposite trends that both population density and POI density are negatively correlated with the incidence rate. Similar but much more remarkable downtrends can be observed for the incidence rate of obesity, in which high intensity plays the most interactive role and shows a consistent changing trend with primary features.

When it comes to mental health, the overall trend toward open space is not obvious. In the range of a low proportion of open space, the predicted proportion of mental health problems drops sharply, then tends to be stable, and rise slightly in the range of high values. The effect of population density shows an overall trend of rising first and then declining. With the increase in population density, the proportion of mental health problems predicted increases, while after the population density increases to a certain value, the proportion of mental health problems predicted decreases, and the downward trend gradually slows down. According to the color distribution of SHAP dependence contribution plots, the most significant factor interacting with the open space is high intensity. The closer the color is to blue, the greater the proportion of high intensity development. The color distribution of points with different y values in the same x interval is analyzed and compared. In the low open space interval, the proportion of high-density development changes most sharply from high to low, which suggests that the combination of the high-density development ratio and the open space ratio leads to a sharp decline in the prediction of mental health problems. The most significant factor that interacts with population density is also high intensity. With the lower population density range, the higher proportion of high intensity development corresponds to the higher predicted value, while the change trend in the high population density range is not significant.

As for lack of sleep, the degree of crowding has a positive relationship with the ratio of lack of sleep, but the growth of open space is related to the decline of lack of sleep. The POI density presents a sharp downtrend at the low density range and a gradually rising trend

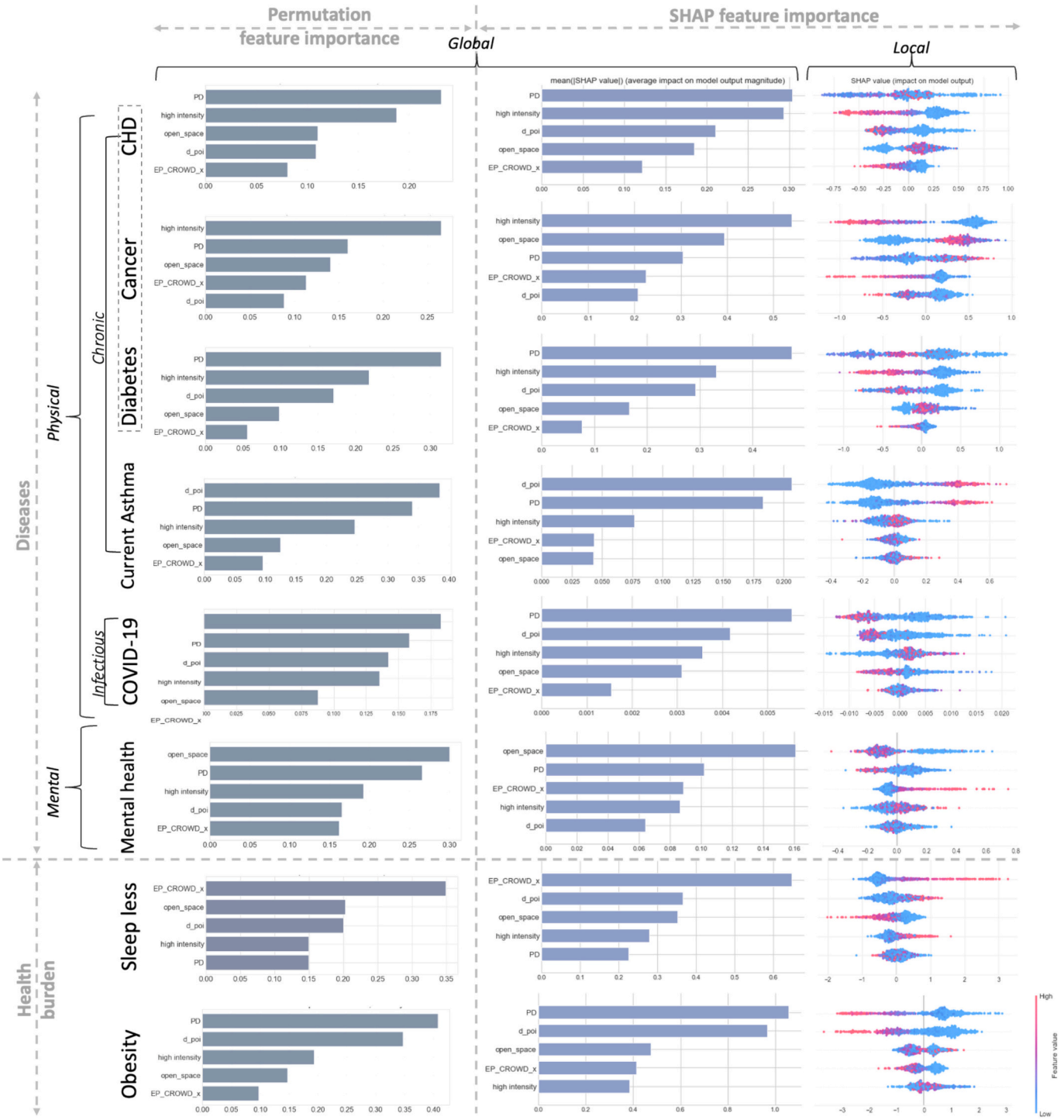


Fig. 4. Results of feature importance analysis.

thereafter. By checking the most interactive feature, we might deduce that the open space acted as the leading factor at the low range of POI density (Fig. 6).

5. Discussion

In this study, we have applied interpretable machine learning methods to urban areas, establishing effective techniques that expand the potential applications within urban studies. This research is among the early efforts to amalgamate multiple health risks, encompassing chronic and infectious diseases, health burdens, and mental health. We have demonstrated the significant associations between urban density and public health. Then we provide evidence that the correlations between urban density and health risks are complex and non-linear, consistent with the previous research (Stevenson & Gleeson, 2019). In particular, the approaches of permutation feature importance, PDP, ICE, and Shapley values show more details from global and local views of key influencing features, non-linear trends, key thresholds, and interactions of features.

Compared to other statistical methods and traditional machine learning models, interpretable machine learning exhibits clear advantages in transparency and comprehensibility of the model's decision-

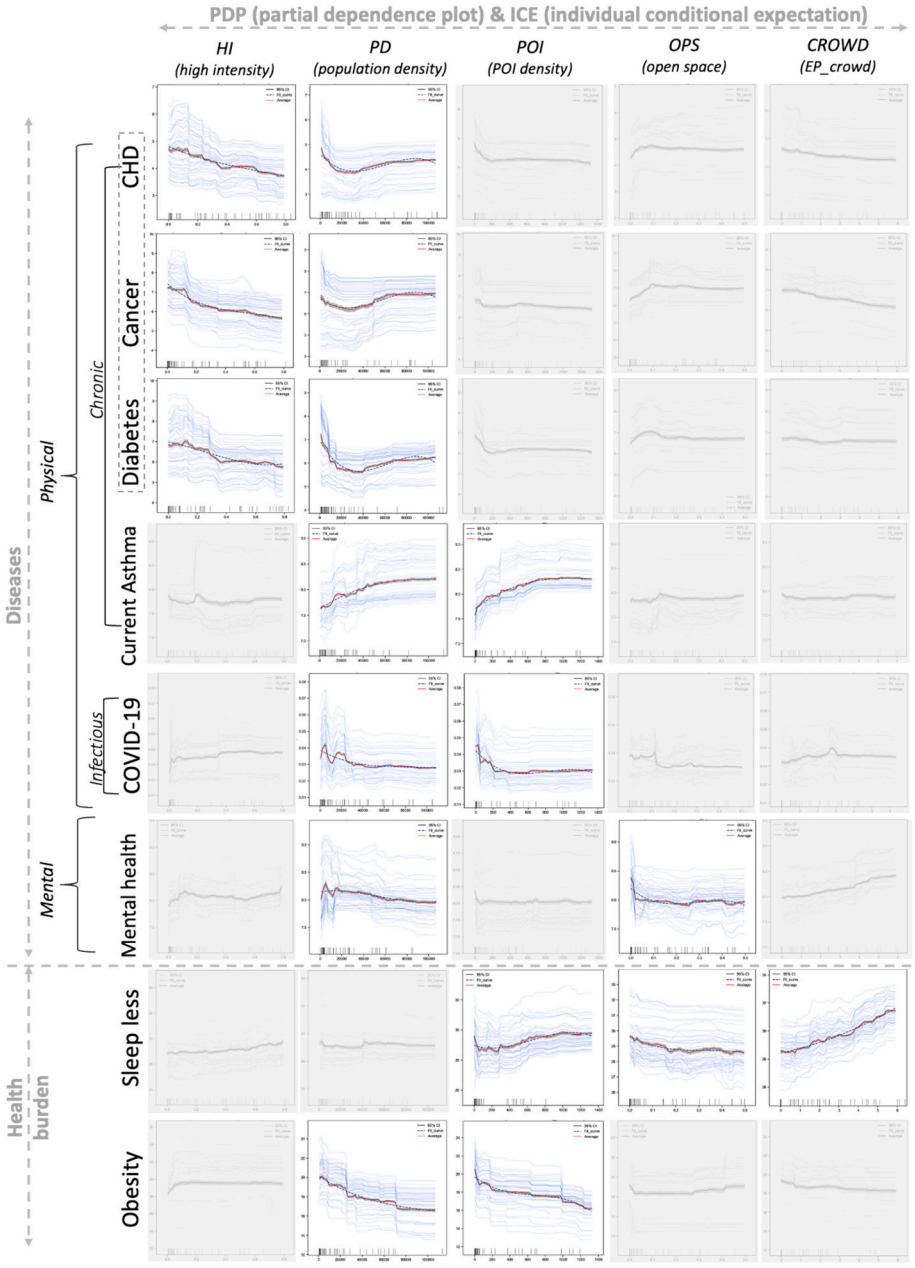


Fig. 5. Results of PDP and ICE.

Table 5
Results of random forest models.

RF	MAE	MSE	RMSE	R2
Obesity	0.52	0.46	0.68	0.54
Sleep < 7	0.70	0.81	0.90	0.16
Current asthma	0.61	0.63	0.79	0.32
Cancer	0.53	0.52	0.72	0.47
Diabetes	0.57	0.61	0.78	0.44
CHD	0.55	0.56	0.75	0.43
COVID19	0.02	0.001	0.04	0.04
Mental health	0.70	0.82	0.90	0.07

Table 6
Results of XGBoost models.

XGBoost	MAE	MSE	RMSE	R2
Obesity	0.56	0.52	0.72	0.48
Sleep < 7	0.76	0.92	0.96	0.05
Current asthma	0.65	0.73	0.86	0.21
Cancer	0.56	0.54	0.73	0.45
Diabetes	0.62	0.66	0.81	0.39
CHD	0.61	0.59	0.77	0.39
COVID19	0.02	0.001	0.04	0.02
Mental health	0.77	0.96	0.99	–

making process. Firstly, interpretable machine learning methods like feature importance help us efficiently filter the key factors. Even though we do not include an extensive range of urban density indicators in this research, sorting features by their importance helped us eliminate redundancy. Additionally, interpretable machine learning methods such as PDP, ICE, and SHAP play crucial roles in delineating non-linear and complex relationships. Numerous health-related researchers have highlighted the complexity of influencing systems, which are typically non-linear (Ahmad et al., 2018; Rudin, 2019; Stiglic et al., 2020). By conducting appropriate interpretable machine learning methods, the

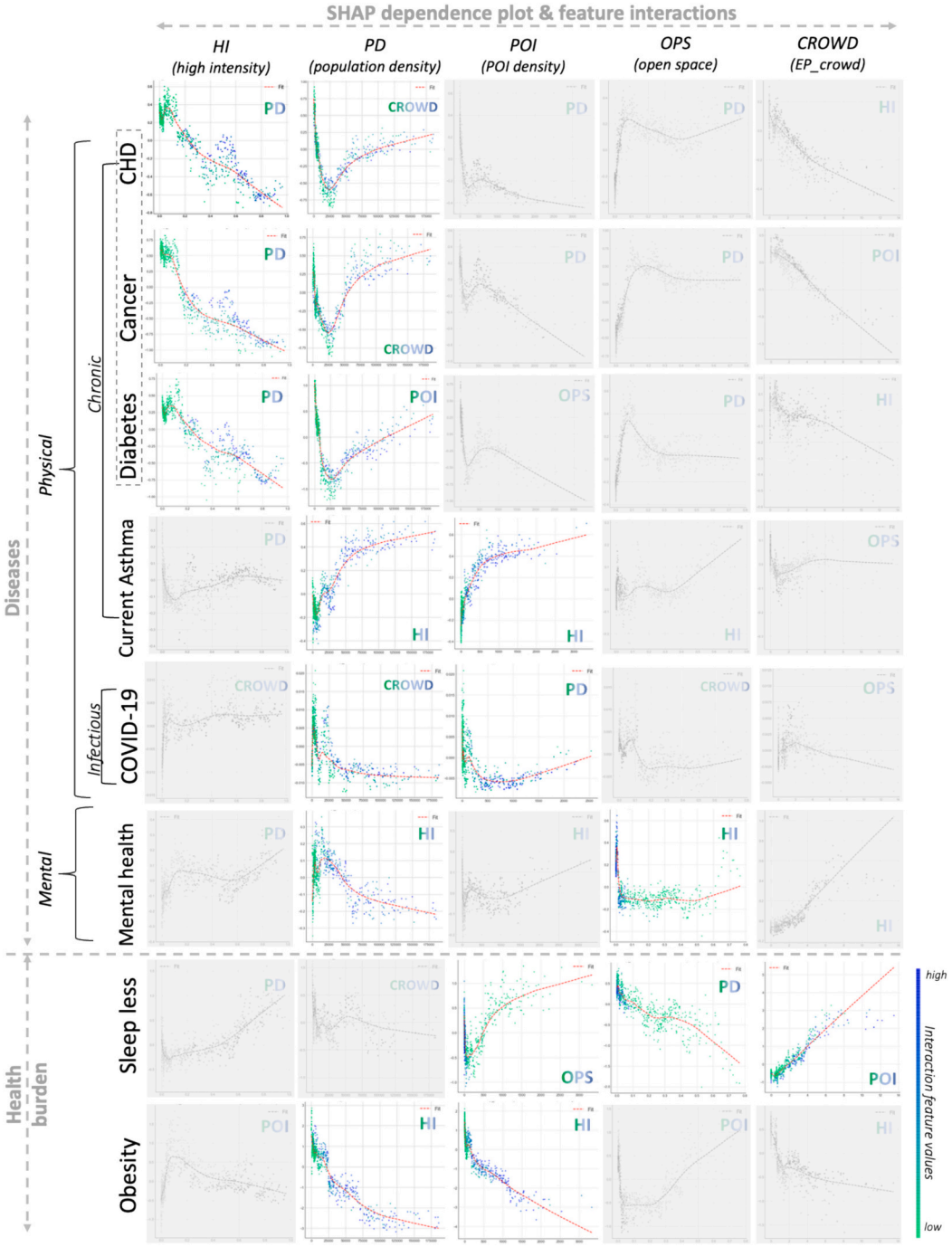


Fig. 6. Results of SHAP dependence plots.

non-linear relationships, which couldn't be deduced by general statistical methods or explained by traditional machine learning methods, could be performed vividly and clearly. Increasingly, researchers acknowledge the limitations of causal theory, particularly in fields such as public health and urban studies (Batty, 2016; Stevenson & Gleeson, 2019). The SHAP method in interpretable machine learning adds more detail to feature interactions, revealing nuanced rules and mechanisms beyond direct and simple causation.

Noteworthy, we find some general and special characteristics of multiple health risks, supporting and supplementing other research. As for physical diseases, long-term chronic diseases like CHD, cancer, and diabetes show high similarities. The factors of population density and high intensity play important roles in influencing the incidence rates. With the increase in population density, the incidence rates tend to decrease first and then increase. And high intensity performs a negative correlation with the incidence rates. Many previous studies find negative relationships between population density and typical chronic diseases, such as heart disease analysis in Japan and the Metropolitan statistical area (Griffin et al., 2013; Konishi et al., 2020), diabetes in Toronto (Glazier et al., 2014). Some studies reveal the increasing trends of chronic diseases in densely populated areas, like cancer in Taiwan (Yang & Hsieh, 1998), cardiovascular disease in Wuhan (Li et al., 2022), lung cancer in Shanghai (Wang et al., 2022). There are many reasons causing this variance, among which the choice of scales and regions might play the leading roles. The decreasing trends between chronic diseases and population density are mainly detected at larger scales in

relatively low-dense areas. While the positive relationships between chronic diseases and population density are generally in relatively dense areas. The combination of these two low- and high-density studies strongly implies the reliability of our findings. Few have directly conclude on the high intensity. Short-term chronic conditions, such as current asthma, and infectious diseases like COVID-19, are more influenced by factors like population density and POI density. Current asthma presents negative correlations with density factors, and COVID-19 shows positive correlations. But they both display a reduction in slopes as the density got higher. In line with these, lots of evidence from COVID-19's spread shows that denser areas tend to have early and rapid outbreaks (Yip et al., 2021; Li, Peng, et al., 2021; Hu et al., 2021; Zhang, 2020; Mollalo et al., 2020), which is consistent with our findings. A similar study done in LA also shows an increased population density is associated with a decrease in health problem (Kim et al., 2021). Moreover, changes in high-density ranges should not be overlooked, indicating a super-linear relationship between density and infectious diseases to some extent. As for the health burden, obesity is negatively correlated with population density and POI density. A large number of studies on obesity have proven that (Ewing et al., 2003; Lopez, 2007; Rundle et al., 2007). In contrast, lack of sleep is positively related to POI density and the degree of crowding, and the expansion of open space would lead to a decrease in lack of sleep. Indirect evidence that the likelihood of short sleep becomes lower with more tree canopy and less noise exposure has been proven in previous research. The relationships shown in mental health are not obvious. Both population density and open space present great effects in the low ranges but smaller fluctuations in the rest. Although the results of SHAP dependence and interaction analysis provide more evidence on feature impacts, suggesting that the interactions among different factors might cause the various trends for the multiple ranges, the current results are not convincing enough to figure out the leading elements.

6. Conclusion

In this paper, a ML-based approach for understanding non-linear and complex associations has been developed and validated in cases of American metropolitan areas. Several interpretable methods are conducted to explain the decision-making process of ML models. By comprehensively studying multiple health risks and their associations with urban density, some valuable findings are obtained from this research to help answer the questions of the relationships and mechanisms between urban density and health risks.

Our work could contribute convincing evidence to at least two kinds of opinions. Firstly, we prove that the associations between urban density and health risks were not a by-product of socioeconomics. Urban density itself does significantly influence human health. Secondly, we find that the extent of interaction between health issues and the environment could be an essential clue in distinguishing patterns of relationships. The increase in the extent of interactions might lead to the negative effects of density indicators. More specifically, for different types of diseases, long-term chronic diseases tend to have negative relationships with density factors in which the proportion of high-intensity developed space has greater roles, compared to short-term infectious diseases like health burden and mental health problems that tend to present positive relationships with urban density, in which population and activity-related density factors dominate. For the same type of diseases, such as chronic diseases, long-term and more-accumulating diseases of CHD, diabetes, and cancer, tend to show more negative trends as density increased compared to short-term and less-accumulating diseases of current asthma, and high intensity gives its way of affecting the outcomes of POI density in this process. In addition, the relationships were non-linear, in which the factor of population density indicated obvious tuning points. Though the interactions between features could explain the fluctuations to some degree, we could not confirm whether they resulted from reality or the model's misunderstanding.

Apparently, the findings discussed here are only the tip of the iceberg. There are several limits in this study. Firstly, though we define the urban density from perspectives of population, built environment and activity, we miss some key indicators like employment density, land use diversity, road networks, and social tie density due to the data availability. Secondly, samples are limited to the metropolitan areas of the United States, and data are just collected from several open sources. More studies are needed to compare the regional differences, include dynamic factors, and investigate wide-range health problems. Besides, the mechanism behind it deserves deeper thinking in human behaviors, biology, medical science, and other fields. Future research may benefit from the design and implementation of clinical experiments.

CRediT authorship contribution statement

Zerun Liu: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Chao Liu:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgement

Peking University Lincoln Institute Funds 2021-2022 (FS01-20211215-LC)

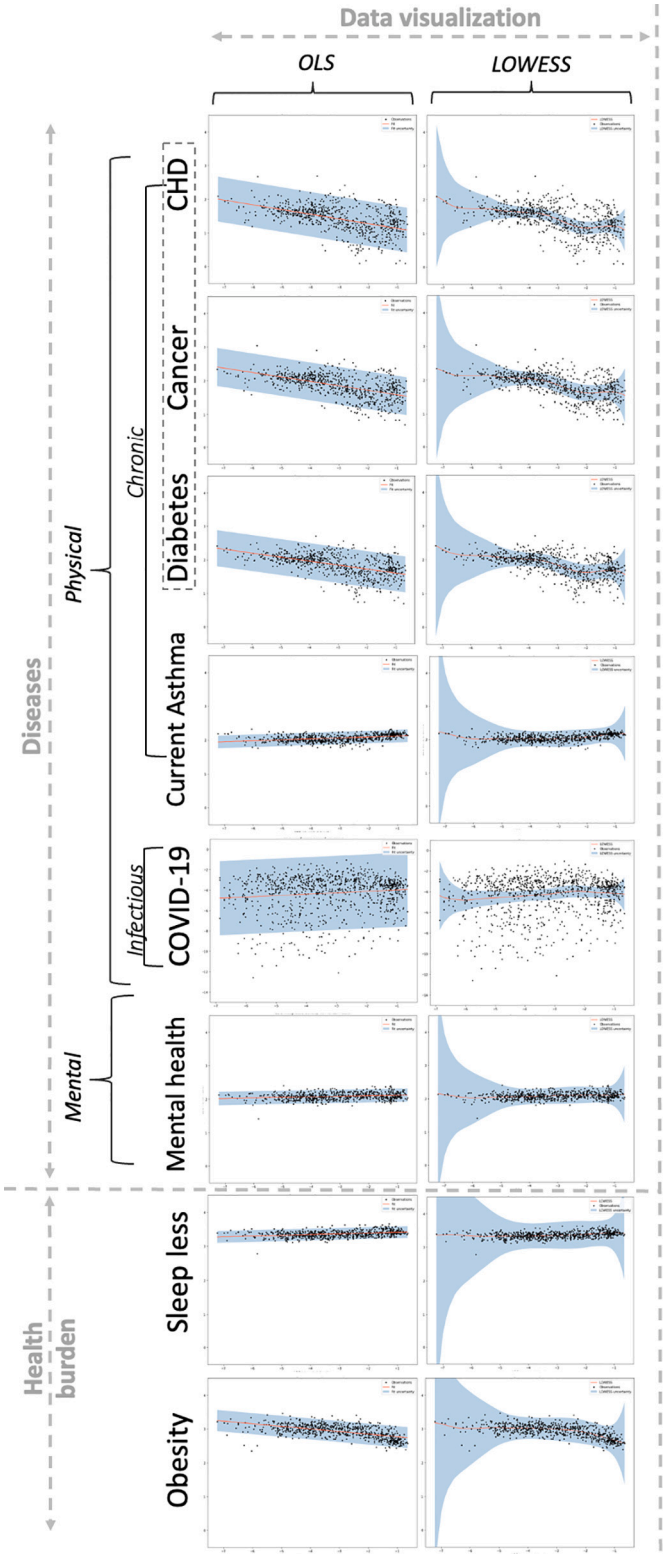
Ministry of Housing and Urban-Rural Development Research Project 2021-2022 (2021-k-008) Shanghai Rising-Star Program 22QB1404900

Chao Liu

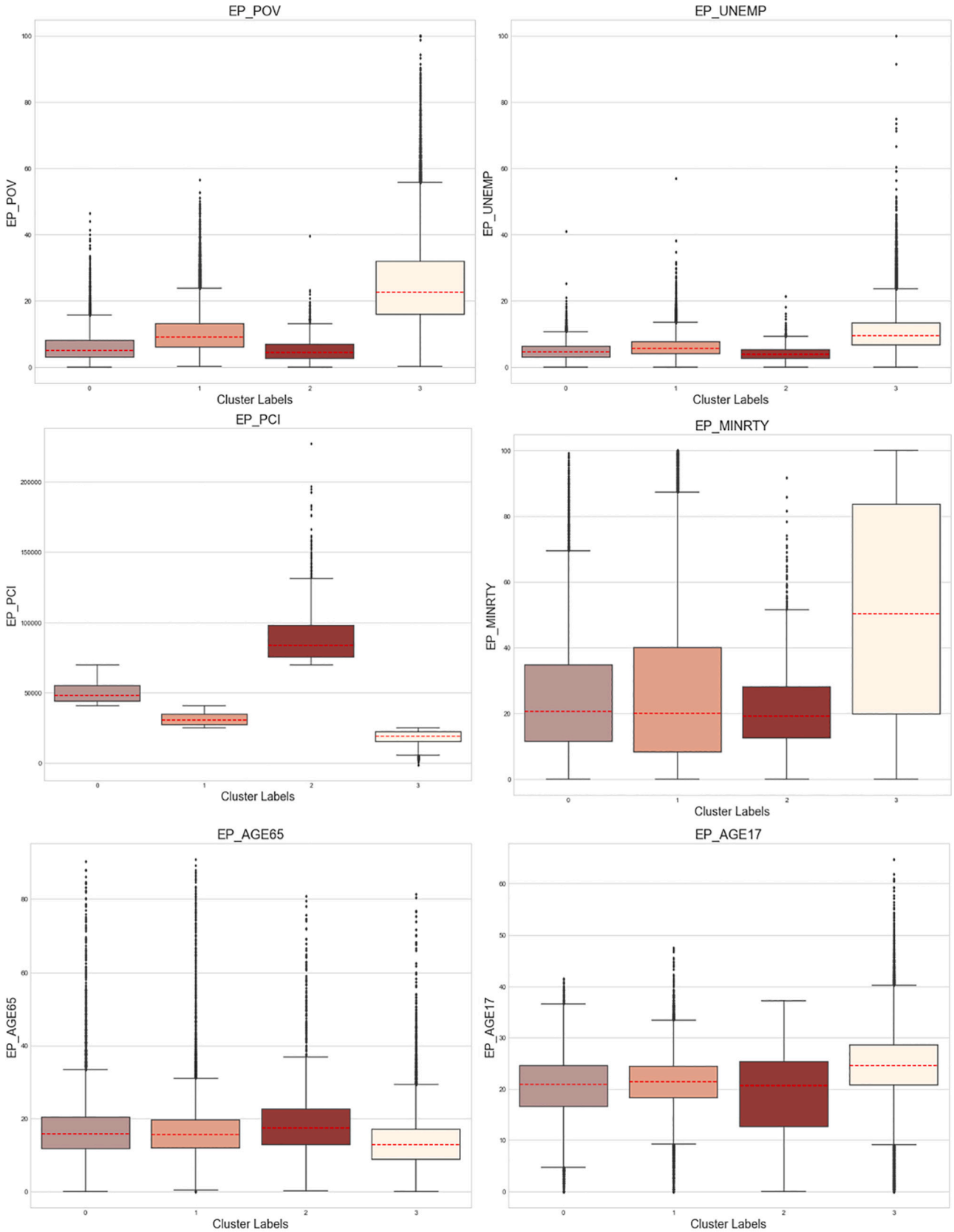
Appendix 1. Data visualization

We started this analysis by simply visualizing the relationships between multiple health risks and the integrated density indicators. Since the distributions of integrated density indicators and incidence rates were right-skewed, we used log transformation to make them conform to normality, which could help us observe patterns in relationships. Thus, the x-axis represented the log-transformed integrated density indicators, and the y-axis represented the log-transformed incidence rates of multiple health risks. Then we plotted linear fit (OLS methods) and curve fit (LOWESS methods) to basically abstract the patterns shown in the visualization. Almost all the data points in the sample were located within the 95 % confidence interval, indicating the reliability of our fitting. Specific to each type of health risks, for chronic diseases such as CHD, diabetes, and cancer showed consistent trends of change, that is, as the integrated density indicators increased, the incidence rate gradually decreased. And the curve fit showed that when the density increased to a certain level, the downward trends in diseases slowed down. However, the current asthma performed the opposite trends. The

increase in integrated density indicators led to a rise in incidence rates. As for infectious disease, the infection rates of COVID-19 were positively correlated with density. And for the health burden, the rates of obesity had a negative relationship with the integrated density indicators. The lack of sleep showed a positive relationship, but this tendency was not significant. Similarly, mental health increased with the increase in integrated density, and the linear growth trend was also very moderate. Though the linear fit in our log-transformed data visualization demonstrated a power-law relation, in which the slope was equal to the exponent, the curve fit implied that the data did not obey the perfect power-law relation and performed certain non-linear complex relationships, especially in the high density interval.



Appendix Fig. 1. Data visualizations of multiple health risks and integrated density indicators.



Appendix Fig. 2. Analysis of clustering.
Boxplots of socioeconomic factors for clusters

References

Ahmad, M. A., Eckert, C., & Teredesai, A. (2018). Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (pp. 559–560).

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics, 26*, 1340–1347.

Bassolas, A., Barbosa-Filho, H., Dickinson, B., Dotiwalla, X., Eastham, P., Gallotti, R., … Ramasco, J. J. (2019). Hierarchical organization of urban mobility and its connection with city livability. *Nature Communications, 10*, 4817.

Batty, M. (2016). Complexity in city systems: Understanding, evolution, and design. In *A planner's encounter with complexity*. Routledge.

Becker, G. S., Glaeser, E. L., & Kevin, M. (1999). Population and economic growth. *J American Economic Review Murphy, 89*, 145–149.

Bettencourt, L. M. A., Lobo, J., Helbing, D., Kühnert, C., & West, G. B. (2007). Growth, innovation, scaling, and the pace of life in cities. *104*, 7301–7306.

Bettencourt, L., & West, G. (2010). A unified theory of urban living. *Nature, 467*, 912–913.

Bi, Y., Xiang, D., Ge, Z., Li, F., Jia, C., & Song, J. (2020). An interpretable prediction model for identifying N7-methylguanosine sites based on XGBoost and SHAP. *Molecular Therapy–Nucleic Acids, 22*, 362–372.

Carlino, G. A., Chatterjee, S., & Hunt, R. M. (2007). Urban density and the rate of invention. *Journal of Urban Economics, 61*, 389–419.

Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Division of Population Health. (2016-2022). In Centers for Disease Control and Prevention (Ed.), *PLACES: Local data for better health, place data*.

Chan, S. M., Wong, H., Chung, R. Y.-N., & Au-Yeung, T. C. (2021). Association of living density with anxiety and stress: A cross-sectional population study in Hong Kong. *Health & Social Care in the Community, 29*, 1019–1029.

Philippa Clarke, University of Michigan. Institute for Social Research; Robert Melendez, University of Michigan. Institute for Social Research. 2001–2016. "National Neighborhood Data Archive (NaNDA): Land cover by census tract, United States." In.

Dong, E., Du, H., & Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases, 20*(5), 533–534.

Dovey, K., & Pafka, E. (2014). The urban density assemblage: Modelling multiple measures. *Urban Design International, 19*, 66–76.

Ewing, R., Schmid, T., Killingsworth, R., Zlot, A., & Raudenbush, S. (2003). Relationship between urban sprawl and physical activity, obesity, and morbidity. *American Journal of Health Promotion, 18*, 47–57.

Foundation: Cambridge, UK. (2020). In OpenStreetMap contributors (Ed.), *OpenStreetMap database [PostgreSQL via API]*.

Geospatial Research, Analysis, and Services Program (GRASP). (2021). *CDC/ATSDR SVI data*.

Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. *Journal of Economic Geography, 1*, 27–50.

Glazier, R. H., Creatore, M. I., Weyman, J. T., Fazli, G., Matheson, F. I., Gozdyra, P., … Booth, G. L. (2014). Density, destinations or both? A comparison of measures of walkability in relation to transportation behaviors, obesity and diabetes in Toronto, Canada. *PLoS One, 9*, Article e85295.

González, M. C., Hidalgo, C. A., & Barabási, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*, 779–782.

Griffin, B. A., Eibner, C., Bird, C. E., Jewell, A., Margolis, K., Shih, R., … Escarce, J. J. (2013). The relationship between urban sprawl and coronary heart disease in women. *Health & Place, 20*, 51–61.

Gruebner, O., Rapp, M. A., Adli, M., Kluge, U., Galea, S., & Heinz, A. (2017). Cities and mental health. *Deutsches Ärzteblatt International, 114*, 121.

Hamidi, S., Sabouri, S., & Ewing, R. (2020). Does density aggravate the COVID-19 pandemic? *Journal of the American Planning Association, 86*, 495–509.

Hipp, J. A., & Chalise, N. (2015). Peer reviewed: Spatial analysis and correlates of county-level diabetes prevalence, 2009–2010. *Preventing Chronic Disease, 12*.

Howard, E. (1902). *Garden city of tomorrow*. London: Passim.

Hu, M., Roberts, J. D., Azevedo, G. P., & Milner, D. (2021). The role of built and social environmental factors in covid-19 transmission: A look at america's capital city. *Sustainable Cities and Society, 65*.

Kim, Y., Cho, J., Wen, F., & Choi, S. (2021). The built environment and asthma: Los Angeles case study. *Journal of Public Health*, 1–8.

Konishi, M., Matsuzawa, Y., Ebina, T., Kosuge, M., Gohbara, M., Nishimura, K., … Tsutsui, H. (2020). Impact of population density on mortality in patients hospitalized for heart failure–JROAD-DPC Registry Analysis. *Journal of Cardiology, 75*, 447–453.

Levinson, D. (2012). Network structure and city size. *PLoS One, 7*, Article e29721.

Li, B., Peng, Y., He, H., Wang, M., & Feng, T. (2021). Built environment and early infection of COVID-19 in urban districts: A case study of Huangzhou. *Sustainable Cities and Society, 66*, Article 102685.

Li, S., Ma, S., & Zhang, J. (2021). Association of built environment attributes with the spread of COVID-19 at its initial stage in China. *Sustainable Cities and Society, 67*, Article 102752.

Li, X., Zhou, L., Liu, X., Qianqian Dun, L., & Ma, and Yuliang Zou.. (2022). Community built environment and the associated ischemic heart disease risk: Evidence from multi-source data in Wuhan, China. *Journal of Transport & Health, 25*, Article 101371.

Lopez, R. P. (2007). Neighborhood risk factors for obesity. *Obesity, 15*, 2111–2119.

Louf, R., & Barthelemy, M. (2014). How congestion shapes cities: From mobility patterns to scaling. *Scientific Reports, 4*, 5561.

Lu, Y., Xiao, Y., & Ye, Y. (2017). Urban density, diversity and design: Is more always better for walking? A study from Hong Kong. *Preventive Medicine, 103*, S99–S103.

Melis, G., Gelormino, E., Marra, G., Ferracin, E., & Costa, G. (2015). The effects of the urban built environment on mental health: A cohort study in a large northern Italian City. *International Journal of Environmental Research and Public Health, 12*, 14898–14915.

Milgram, S. (1974). The experience of living in cities. In W. J. H. Sims, & D. D. Bauman (Eds.), *Human behavior, the environment: Interactions between man, and his physical world* (pp. 217–240).

Mollalo, A., Vahedi, B., & Rivera, K. M. (2020). GIS-based spatial modeling of COVID-19 incidence rate in the continental United States. *Sci Total Environ, 728*, Article 138884.

Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.).

Pan, W., Ghoshal, G., Krumme, C., Cebrian, M., & Pentland, A. (2013). Urban characteristics attributable to density-driven tie formation. *Nature Communications, 4*, 1961.

Rader, B., Scarpino, S. V., Nande, A., Hill, A. L., Adlam, B., Reiner, R. C., … Kraemer, M. U. G. (2020). Crowding and the shape of COVID-19 epidemics. *Nature Medicine, 26*, 1829–1834.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*, 206–215.

Ruiz-Tagle, J., & Urria, I. (2022). Household overcrowding trajectories and mental well-being. *Social Science & Medicine, 296*, Article 114051.

Rundle, A., Diez, A. V., Roux, L. M., Freeman, D. M., Neckerman, K. M., & Weiss, C. C. (2007). The urban built environment and obesity in New York City: A multilevel analysis. *American Journal of Health Promotion, 21*, 326–334.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature, 484*, 96–100.

Stevenson, M., & Gleeson, B. (2019). Complex urban systems: Compact cities, transport and health. In *Integrating human health into urban and transport planning: A framework* (pp. 271–285).

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., & Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10*, Article e1379.

Townshend, T., & Lake, A. A. (2009). Obesogenic urban form: Theory, policy and practice. *Health & Place, 15*, 909–916.

Wang, L., Sun, W., Moudon, A. V., Zhu, Y.-G., Wang, J., Bao, P., … Zhang, S. (2022). Deciphering the impact of urban built environment density on respiratory health using a quasi-cohort analysis of 5495 non-smoking lung cancer cases. *Science of the Total Environment, 850*, Article 158014.

WHO. (2009). *Global health risks: Mortality and burden of disease attributable to selected major risks*. World Health Organization.

Xie, J., & Zhu, Y. (2020). Association between ambient temperature and COVID-19 infection in 122 cities from China. *Science of the Total Environment, 724*, Article 138201.

Xu, F., Li, J. Q., Liang, Y. Q., Wang, Z. Y., Hong, X., Ware, R. S., … Owen, N. (2010). Residential density and adolescent overweight in a rapidly urbanising region of mainland China. *Journal of Epidemiology & Community Health, 64*, 1017–1021.

Yang, C.-Y., & Hsieh, Y.-L. (1998). The relationship between population density and cancer mortality in Taiwan. *Japanese Journal of Cancer Research, 89*, 355–360.

Yin, C., Yao, X., & Sun, B. (2022). Population density and obesity in rural China: Mediation effects of car ownership. *Transportation Research Part D: Transport and Environment, 105*, Article 103228.

Yip, T. L., Huang, Y., & Liang, C. (2021). Built environment and the metropolitan pandemic: Analysis of the COVID-19 spread in Hong Kong. *Building and Environment, 188*, Article 107471.

Zhang, W. (2020). The impact of the built environment on the COVID-19 epidemic and evidence-based practice: A preliminary analysis of the distribution of COVID-19 in American cities. *City Planning Review, 44*, 33–41.