JAMIA Open, 3(3), 2020, 386–394
doi: 10.1093/jamiaopen/ooaa038
Advance Access Publication Date: 28 October 2020
Research and Applications

Research and Applications

Using electronic health records for population health sciences: a case study to evaluate the associations between changes in left ventricular ejection fraction and the built environment

Yiye Zhang,[1,2] Mohammad Tayarani,[3] Subhi J. Al'Aref,[4] Ashley N. Beecy,[5] Yifan Liu,[1] Evan Sholle,[1] Arindam RoyChoudhury,[1] Kelly M. Axsom,[6] Huaizhu OliverGao,[3] Jyotishman Pathak,[1] and Jessica S. Ancker[1]

[1]Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, New York City, New York, USA, [2]Department of Emergency Medicine, Weill Cornell Medicine, Cornell University, New York City, New York, USA, [3]School of Civil and Environmental Engineering, Cornell University, Ithaca, New York, USA, [4]Division of Cardiology, Department of Medicine, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA, [5]Division of Cardiology, Department of Medicine, Weill Cornell Medicine, New York, New York, USA and [6]Columbia University Irving Medical Center, New York, New York, USA

Corresponding Author: Yiye Zhang, PhD, MS, Department of Population Health Sciences, Weill Cornell Medicine, Cornell University, 425 East 61st Street, New York, NY 10065, USA; yiz2014@med.cornell.edu

Received 12 March 2020; Revised 16 July 2020; Editorial Decision 10 August 2020; Accepted 20 August 2020

ABSTRACT

Objective: Electronic health record (EHR) data linked with address-based metrics using geographic information systems (GIS) are emerging data sources in population health studies. This study examined this approach through a case study on the associations between changes in ejection fraction (EF) and the built environment among heart failure (HF) patients.
Materials and Methods: We identified 1287 HF patients with at least 2 left ventricular EF measurements that are minimally 1 year apart. EHR data were obtained at an academic medical center in New York for patients who visited between 2012 and 2017. Longitudinal clinical information was linked with address-based built environment metrics related to transportation, air quality, land use, and accessibility by GIS. The primary outcome is the increase in the severity of EF categories. Statistical analyses were performed using mixed-effects models, including a subgroup analysis of patients who initially had normal EF measurements.
Results: Previously reported effects from the built environment among HF patients were identified. Increased daily nitrogen dioxide concentration was associated with the outcome while controlling for known HF risk factors including sex, comorbidities, and medication usage. In the subgroup analysis, the outcome was significantly associated with decreased distance to subway stops and increased distance to parks.
Conclusions: Population health studies using EHR data may drive efficient hypothesis generation and enable novel information technology-based interventions. The availability of more precise outcome measurements and home locations, and frequent collection of individual-level social determinants of health may further drive the use of EHR data in population health studies.

Key words: cardiovascular diseases, built environment, public health informatics, geographic information system, electronic health records

LAY SUMMARY

Electronic health record (EHR) data linked with address-based metrics using geographic information systems (GIS) are emerging data sources in population health studies. We examined the relationship between the built environment and the changes in ejection fraction (EF) using EHR data at an urban academic medical center. Longitudinal clinical information of 1287 heart failure (HF) patients was linked with address-based built environment metrics related to transportation, air quality, land use, and accessibility by GIS. The primary outcome is the increase in the severity of EF categories. Statistical analyses were performed using mixed-effects models, including a subgroup analysis of patients who initially had normal EF measurements. Increased daily nitrogen dioxide concentration was associated with the outcome while controlling for known HF risk factors including sex, comorbidities, and medication usage. In the subgroup analysis that was performed among patients with initially normal EF measurements, the outcome was significantly associated with decreased distance to subway stops and increased distance to parks. The found association on air quality is consistent with known literature, whereas the accessibility to the subway and parks present new evidence to be validated with bigger datasets in the future. EHR data may drive efficient hypothesis generation for population health studies.

INTRODUCTION

Population health studies have commonly been defined by cohort identification and follow-up in the last decades.[1] The success of population health studies is largely determined by the available funding to define and follow-up patient cohorts as well as the theories and hypotheses that drive the study designs. Today, researchers across the domains in medicine and healthcare are increasingly drawn to electronic health records (EHRs), a source of routinely collected observational health data that potentially reduces the burden of cohort identification and follow-up while accelerating the hypothesis generation process. EHR data have seen a wide variety of use cases, ranging from disease prediction, computational phenotyping, drug discovery, to personalized treatment strategies.[2–6] Cited for its routine availability, large volume, and rich details, EHR data are expected to drive automation and innovation for operational tasks and research studies that are traditionally knowledge- and labor-intensive.

In this study, we aimed to evaluate the feasibility of using EHR data in population health studies to examine the associations of clinical outcomes with environmental exposures.[7,8] In particular, this case study focused on the built environment, which refers to the human-made environment through urban planning, such as buildings to provide food and shelter, infrastructure for public transportation, and space for social activities.[9] The built environment is considered to influence public health through several mechanisms, including air quality, noise level, and access to healthy lifestyles. Notably, significant health effects by the built environment on heart failure (HF) patients have been reported by previous research.[9] HF is among the leading causes of morbidity, mortality, and substantial healthcare expenditure in the United States.[10] Its global prevalence is estimated to be more than 26 million, a figure projected to increase further as the global population continues to age.[10]

Previous population health and epidemiologic studies have identified risk factors of HF incidence and mortality, including male sex, high blood pressure, coronary artery disease, diabetes, valvular heart disease, tobacco use, obesity, low education level, and socioeconomic deprivation.[11] Among the built environment factors, the effects from air quality and roadway proximity have been reported in multiple studies. HF incidence has been associated with exposure to particulate matter $\leq$2.5 μm in aerodynamic diameter ($PM_{2.5}$) in a 4-year prospective cohort study of women across the United States,[12] and an 11.5-year prospective cohort study in Europe.[13] HF mortality was associated with exposure to $PM_{2.5}$ in the Cancer Prevention Study II of 1.2 million adults over a 16-year follow-up.[14] HF mortality was also associated with roadway proximity and noise volume in 5-year follow-up studies in Worcester, Massachusetts, a 9-year cohort study in the Netherlands, and a cross-sectional survey in Toronto, Canada, respectively.[15–17]

Observational data such as national Medicare claims and registries have been used in studies to identify associations between HF hospital admission rates and air pollutants,[8,19] and between socioeconomic deprivation and the cardiovascular events.[20] Compared to claims data, EHRs contain richer clinical information from structured and unstructured laboratory and imaging test results without requiring the upfront investment of creating registries. Through geocoding of patient residential location information in EHRs and further linking it to publicly available environmental data sources, recent studies have identified associations such as air pollution and cardiovascular events during labor and delivery,[21] air pollution and asthma,[7] among others.[8]

Limitation of using EHR data in observational studies has been discussed in previous and recent literature.[22,23] Particularly, it is known that EHR-derived cohorts potentially lead to erroneous and biased association estimates due to incomplete data collection and censoring especially in fragmented healthcare markets. Nevertheless, evaluating the ability to detect known population health associations in the EHR data may pave the path for more rigorous data collection efforts using the EHR, potentially starting to address current limitations.[24,25] Thus, we sought to contribute to the existing literature on EHR-based population health studies by conducting a case study of whether previously reported effects of the built environment on HF patients' cardiovascular functions could be identified from the EHR data combined with address-based metrics. Furthermore, we explored whether additional associations would be found using EHR data for future hypothesis-generating studies.

MATERIALS AND METHODS

Study setting

This study was performed at an academic medical center in a dense, urban environment in New York City. Study data were extracted from a commercial EHR and transformed to the Observational Medical Outcomes Partnership (OMOP) common data model maintained by the Observational Health Data Sciences and Informatics consortium.[26] Weill Cornell Medicine Internal Review Board

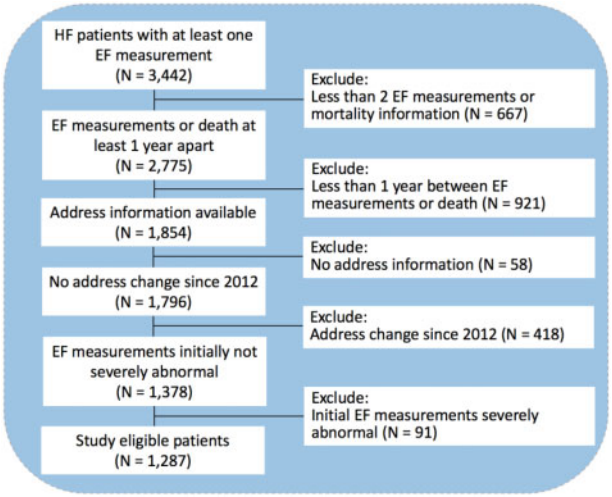

Figure 1. Patient inclusion and exclusion criteria.

approved the study design and its use of protected health information (Protocol#: 1711018789).

Participants

We included patients if they had an encounter at the studied medical center between 2012 and 2017 and had a primary or secondary diagnosis of acute or chronic HF. A diagnosis of HF was defined as ICD-9-CM: 428.* or ICD-10-CM: I50*. Patients must have had at least 2 transthoracic echocardiograms with EF measurement that were more than 1 year apart, or have died more than 1 year after the baseline EF measurements. Patients were excluded if they died within 1 year of the baseline EF measurement, if their addresses were not recorded in the EHR, or if address changes were recorded from 2012 to 2017. Lastly, due to the reduced levels of physical activity and subsequent lower exposure to the built environment, patients with severely abnormal EF measurements as defined in the "Outcome Measurement" section at baseline were excluded. The inclusion and exclusion criteria are described in Figure 1.

EHR data

Data elements extracted from patients' EHR data were age, sex, race, average body mass index (BMI), EF, binary indicators for whether patients have ever smoked, binary indicators for whether patients have received at least one prescription of beta blockers (carvedilol, metoprolol, bisoprolol), or renin-angiotensin inhibitors (angiotensin-converting enzyme inhibitor or angiotensin receptor blocker), binary indicators for whether patients have had at least one diagnostic code for hypertension (ICD-9-CM: 401.X-405.X, 437.2 or ICD-10-CM: I10.*, I15.0, I15.8, I67.4), diabetes mellitus (ICD-9-CM: 250.* or ICD-10-CM: E10.*, E11.*), valvular heart disease (ICD-9-CM: 394.*-397.*, V42.2, V43.3 or ICD-10-CM: I05.0, I05.1, I05.2, I05.8, I06.0, I08.0, I08.8, I08.9, I07.1, I07.2, I07.8, Z95.2, Z95.3), coronary artery disease (ICD-9-CM: 410.* -414.*,429.2, V45.81 or ICD-10-CM: I21.09, I21.19, I21.11, I21.29, I21.4, I21.3, I21.9, I21.A1, I21.A9, H18.411, I25.10, I25.2, I20.8, I20.1, I20.8, I20.9, Z95.1), primary care locations, and mortality. Diagnostic codes were extracted from billing diagnoses. Dates of death were obtained through the EHR and the Social Security Death Index. EF measurements were extracted from the unstructured notes of the patients' EHR using a rule-based natural language

Figure 2. Distances to the nearest parks from patients' home locations.

processing method described by Johnson et al.[27] Patients' residential locations were passed to an application programming interface offered by the United States Census Bureau[28] which allowed us to derive both the latitude/longitude pairs and the US census tracts which equal to the 11-digit Federal Information Processing Standard (FIPS) codes.

Public data

The built environment factors on accessibility, traffic, land use, and air quality were extracted at the individual patient level using the aforementioned latitude/longitude pairs in the EHR. In addition, FIPS-level social determinants were extracted based on the 11-digit FIPS code in the EHR.

Four indicators were defined to measure accessibility to public and active transportation and green spaces: the distance to the nearest bus stops, the distance to the nearest subway stops, the distance to the nearest parks, and the distance to the nearest bike facilities. Data on accessibility were obtained from the NYC Department of Planning public data repository.[29] Parks were defined as areas designated as a park, ball field, playground, or public space in NYC Zoning Districts. Figure 2 displays the distances to the park across the 5 boroughs in NYC.

The traffic data were obtained from the New York Best Practice Model, which is an activity-based travel demand model that includes traffic volume on highways, major arterials, and collector's links along with several other transportation measures.[30] The model predicts daily traffic volume in each roadway link for the different types of vehicles including passenger vehicles, buses, taxi, and trucks. We grouped the traffic volumes into 2 groups, respectively, namely, light-duty vehicles such as passenger cars and taxies, and heavy-duty vehicles such as buses and trucks. The stratification controls for the varying environmental impacts by the light- and heavy-duty vehi-

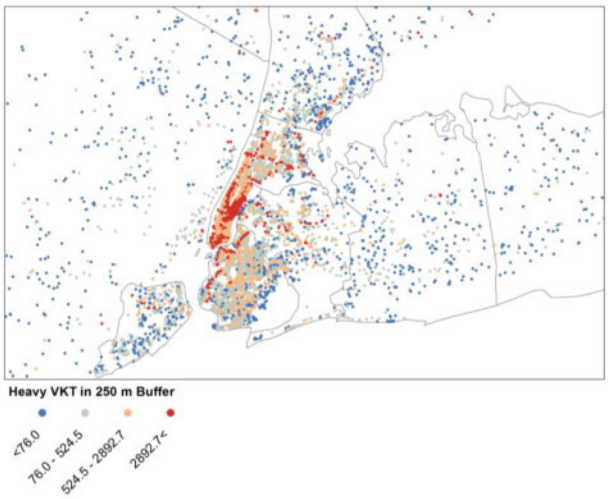

Figure 3. The heavy-duty vehicle activity within 250-m buffer (right) in the studied environment.

cles.[31] Figure 3 displays heavy-duty vehicle activity within 250-m buffers across the 5 boroughs of NYC.

We measured the walkability and availability of a variety of resources for retail, commercial, facility, and residential purposes within 500 m of each patient's home location using the land use mix index and the floor area ratios. The floor area ratio measures the building floor area divided by land area. For example, the areas with a higher share of parking space have lower retail floor area ratio values while areas with smaller setbacks from the street have higher values. Four types of floor area ratios were computed: retail floor area ratio, residential floor area ratio, commercial floor area ratio, and facility floor area ratio.[32] Higher floor area ratios are considered to promote more walkability, an important built environment indicator as our study focuses on an urban environment.[33] Land use data were extracted from the NYC Department of Planning public data repository which includes information about land use type at the parcel level.[29] A measure for the heterogeneity of land use,[34] higher land use mix indices indicates a higher walkability of the area.

For air quality, we estimated patients' exposure to nitrogen dioxide (NO_2) using the Land Use Regression model obtained from the Center for Air, Climate and Energy Solutions.[35] This air pollutant model estimates the daily NO_2 concentration at the block group level using land use regression models and covers both regional and local air pollution hotspots.

Lastly, we obtained census-tract level estimates of social determinants of health including poverty rates, percentages of college degrees, and median home values from the FACETS dataset.[36] While not at the individual-level, these estimates allowed us to control for socioeconomic risk factors identified in previous studies.

Outcome measurement

Left ventricular ejection fraction (EF), the portion of blood pumped out by the left ventricle with each contraction, is one of the most important measurements in diagnosing and defining stages of HF.[37] Under the definitions provided by the American Society of Echocardiography and the European Association of Cardiovascular Imaging,[38] EF measurements are classified into 4 categories: normal (EF >51% in men and EF >53% in women), mildly abnormal (EF between 41%–51% in men and 41%–53% in women), moderately abnormal (EF within 30%–40% in men and women), and severely abnormal (EF <30% in men and women). EF severity may be reflected in HF patients' changing conditions such as shortness of breath and reduced ability for physical activity. EF measurements are most commonly taken with transthoracic echocardiograms and recorded in the EHRs repeatedly following patients' routine care. The outcome in the study is a composite outcome of EF change defined as a deteriorated EF category or mortality within 1 year of a baseline EF measurement. Deteriorated EF category includes a shift from normal to mildly/moderately/severely abnormal, from mildly abnormal to moderately/severely abnormal, or from moderately abnormal to severely abnormal. The composite outcome was not treated as a time-to-event outcome as we assumed that the recorded dates of EF measurements do not equal to the actual time of EF change. Although a prognostic marker such as New York Heart Association (NYHA) Functional Classification would be a strong indicator of HF,[39] this marker was not reliably available in the EHR data, and therefore we defined the study outcomes on the basis of EF measurements.

Statistical methods

Bivariate associations between the exposure and outcome were assessed with chi-squared tests for categorical variables and analysis of variance for continuous variables. Two hypotheses were tested.

> **H1**: The built environment is significantly associated (vs not associated) with a reduction in EF among patients with HF.

Mixed-effects logistic regression with fixed and random effects was used to analyze the associations while controlling for previously reported HF risk factors. HF risk factors that were considered as the fixed effect variables are age, sex, race, BMI, smoking (yes/no), diabetes (yes/no), valvular heart disease (yes/no), coronary artery disease (yes/no), average poverty level at the census tract. The model also contained multiple built environment variables as fixed effects, including floor area ratio for residential use, floor area ratio for facility use, floor area ratio for commercial use, floor area ratio for retail use, land use mix index, average daily NO_2 concentration ($\mu g/m^3$), light-duty vehicle in 250 m buffer in kilometer, heavy-duty vehicle in 250 m buffer in kilometer, distance (km) to nearest bus stops, distance (km) to nearest parks, distance (km) to nearest subway stops, and distance (km) to nearest bike paths. The primary care locations were treated as the random effects in the model to control for the possible care variations across clinics within the health system. Backward elimination was performed for variable selection among the aforementioned variables. Tests for correlations and multicollinearity among variables were tested using the variance inflation factor (VIF). The models were constructed using Stata 14's generalized structural equation model. Since the majority of the patients were age 60 and above, we did not create matched cases and controls by age.

> **H2**: The built environment is significantly associated (vs not associated) with a reduction in EF among HF patients with baseline normal EF.

It is known that the amount of physical activity that may be tolerated by HF patients decreases as the disease progresses. Since reduced physical activity likely leads to different levels of environmental exposure,[40] we examined the effects of the built environment on the outcome among patients whose initial EF measurements were normal as a subgroup analysis.

Sensitivity analyses were conducted. In the first analysis, we strictly used only deterioration in the EF category as the primary outcome of a change in EF. In this analysis, if patients were recorded to have died after at least 1 year following the baseline EF measurement but had no EF measurements that are at least 1 year apart, they were included in the analysis as no-change. Additionally, we performed a second sensitivity analysis limiting the study sample to only patients with both initial and final EF measurements. In the second analysis, if patients were recorded to have died after at least 1 year following the baseline EF measurement but had no EF measurements that are at least 1 year following baseline, they were excluded from the analysis. In addition, on a subset of patients we were able to obtain values of B-type natriuretic peptide (BNP) which is known to be a marker for the severity of acute and chronic HF.[41] BNP values were compared between the group with initial EF measurements studied in the subgroup analysis versus the rest of the study sample whose initial EF measurements were abnormal as defined by the American Society of Echocardiography and the European Association of Cardiovascular Imaging.

RESULTS

A total of 1287 adult patients who met the study criteria were identified. Table 1 lists the variables and their bivariate associations with the outcome. We imputed 215 missing values in BMI using multiple imputation.[42]

Results from the mixed-effects logistic regression for H1 are shown in Table 2. As in previous literature, male sex (OR = 1.093, P value < 0.001) and daily NO_2 concentration (OR = 1.071, P value < 0.001) were significantly associated with increased odds of the outcome. In addition, medication prescription (OR = 1.137, P value < 0.001), age (OR = 0.997, P value < 0.017), BMI (OR = 0.999, P value = 0.001), and Asian race (OR = 0.915, P value < 0.041) were significant in the model. The daily NO_2 concentration remained significant in the sensitivity analyses (see Supplementary Tables SA2 and SA4), and in larger models that included other built environment variables (data not shown). We did not find other risk factors previously reported to be significantly associated. The model had no significant multicollinearity based on VIF (<10).

For H2, a subgroup analysis of the study cohort (N = 1073) whose initial EF was normal is shown in Table 3. Male sex (OR = 1.117, P value < 0.001), BMI (OR = 0.999, P value = 0.037), and medication prescription (OR = 1.158, P value < 0.001) were significantly associated with the outcome. Unlike our findings from the main analysis, increased distance (km) to nearest parks (OR = 1.166, P value = 0.049) and decreased distance (km) to subway stops (OR = 0.947, P value = 0.001) were found to be significantly associated with the outcome. The daily NO_2 concentration was no longer significantly associated with the outcome in the subgroup analysis. Similar results were obtained in the 2 sensitivity analyses (Supplementary Tables SA3 and SA5).

Results from the sensitivity analysis are shown in Supplementary Tables SA1–SA5 in the Appendix for the main cohort and the subgroup analysis cohort. In all analyses, the daily concentration of NO_2 remained significantly associated with the outcome in the main cohort. The associations between the outcome and the distances to parks and subway stations also remained significant in the subgroup analysis cohort. Among the patients who were included in the subgroup analysis, 143 patients had BNP values that were within 2 months of the initial EF measurements used to decide inclusion for subgroup analysis. The distribution of the BNP across EF category is shown in Table 4. The normal group, used for the subgroup analysis, has significantly lower BNP levels compared to the patients whose initial EF measurements were categorized as mild or moderately abnormal (P value = 0.001).

DISCUSSION

The goal of this case study was to identify built environment factors that are associated with a reduction in EF among a cohort of HF patients using EHR data linked with address-based metrics. The daily concentration ($\mu g/m^3$) of NO_2, and accessibility to nearest parks and subway stops, was found in the main and subgroup analysis to be significantly associated with the outcome, respectively. Our finding on the daily concentration ($\mu g/m^3$) of NO_2 agreed with previous studies that examined air quality and cardiovascular events.[12–19] The outcome's association with the distance to the parks in the subgroup analysis has also been reported in previous literature. Given the urban study environment, the associations may be an indicator of increased opportunities in staying physically active. Exercise training has been increasingly reported in recent years to benefit long-term health in HF patients across all ages, gender, and HF severity groups.[43] For example, a recent multicenter randomized clinical trial found exercise training to be associated with modest significant reductions in cardiovascular mortality and HF hospitalization among patients with chronic HF.[44] Specifically related to parks, a randomized crossover study in an urban environment in London, United Kingdom found that walking near a park led to an improvement in lung function, while significant effect of the same exercise was not observed when subjects were walking along a densely populated area.[45] In addition, previous studies have reported that access to parks alleviates stress, improves mental health, and increases subjective well-being, and associated with lower medical expenditures.[46,47] Both stress and poor mental health have well-documented correlative and causative associations with cardiovascular morbidity including HF.[48]

The outcome's association with the distance to nearest subway stops contrasts against previous studies where lack of transportation has been identified as a barrier to healthcare access and thus a risk factor.[49] However, similar to the distance to parks, in the urban setting we studied, our finding may actually reflect the increased likelihood for routine physical activity through walking. As public transportation is by far the most common form of commute in this urban setting, it is possible that longer walks required to reach the subway stops contributed to an increased level of physical activity. It may also explain why the associations from proximity to park and subway stops were only observed in the subgroup analysis since the main analysis included patients with different EF categories and subsequently possible varying levels of physical activity. We aim to explore this association further in future studies. Additionally, while we only studied NO_2 in this study as an indicator of air quality, future studies will also examine the exposure from other air pollutants such as $PM_{2.5}$ in the urban environment.

The use of structured EHR data in our study faced a number of limitations.[22] First, although the healthcare organization had clinics around the city, it is possible that our study data missed EF measurements and other comorbid conditions that were recorded outside the study setting in constructing the models. To address this limitation, we excluded patients who only had 1 EF measurement from our study to better ensure that patients had continuous care within the health system. Additionally, our study data had information on medication but they were limited to prescription and not the actual

Table 1. Descriptive patient characteristics

Variable	Outcome (percentage/standard deviation)	
	No	Yes
Number of patients	887	400
Initial EF category		
Normal	747 (84.22%)	326 (81.50%)
Mildly abnormal	86 (9.70%)	45 (11.25%)
Moderately abnormalse	54 (6.09%)	29 (7.25%)
Last EF category/all-cause mortality*		
Normal	832 (93.80%)	0 (0.00%)
Mildly abnormal	36 (4.06%)	119 (29.75%)
Moderately abnormal	19 (2.14%)	84 (21.00%)
Severely abnormal	0 (0.00%)	81 (20.25%)
All-cause mortality	0 (0.00%)	116 (29.00%)
Sex*		
Female	443 (49.94%)	148 (37.00%)
Male	444 (50.06%)	252 (63.00%)
Race		
Asian	55 (6.20%)	21 (5.25%)
Black or African American	182 (20.52%)	73 (18.25%)
White	321 (36.19%)	150 (37.50%)
Unknown	131 (14.77%)	73 (18.25%)
Other	198 (22.32%)	83 (20.75%)
Age*	68.03 (sd=10.82)	66.44 (sd=12.14)
BMI	29.17 (sd=7.47)	27.88 (sd=6.87)
Smoking (smoker and ex-smoker)		
No	392 (44.19%)	155 (38.75%)
Yes	495 (55.81%)	245 (61.25%)
Valvular heart disease		
No	235 (26.49%)	113 (28.25%)
Yes	652 (73.51%)	287 (71.75%)
Coronary artery disease		
No	89 (10.03%)	34 (8.50%)
Yes	798 (89.97%)	366 (91.50%)
Hypertension		
No	37 (4.17%)	18 (4.5%)
Yes	850 (95.83%)	382 (95.5%)
Diabetes		
No	395 (44.53%)	170 (42.50%)
Yes	492 (55.47%)	230 (57.50%)
Medication*		
No	147 (16.57%)	37 (9.25%)
Yes	740 (83.43%)	363 (90.75%)
Census-tract level poverty rate	18.92% (SD = 0.145)	18.93% (SD = 0.137)
Standardized area for residential use	3.130 (SD = 3.397)	3.373 (SD = 3.591)
Standardized area for commercial use	2.002 (SD = 3.618)	2.111 (SD = 3.711)
Standardized area ratio for retail use	2.132 (SD = 2.754)	2.346 (SD = 3.217)
Standardized land use mix index	8.446 (SD = 10.458)	8.823 (SD = 10.619)
Distance (km) to nearest bus stops	0.103 (SD = 0.117)	0.099 (SD = 0.083)
Distance (km) to nearest subway stops*	0.595 (SD = 0.746)	0.499 (SD = 0.547)
Distance (km) to nearest parks	0.212 (SD = 0.153)	0.222 (SD = 0.163)
Distance (km) to nearest bike paths	0.191 (SD = 0.293)	0.189 (SD = 0.273)
Daily NO_2 concentration ($\mu g/m^3$)*	9.19 (SD = 0.50)	9.27 (SD = 0.51)
Light-duty vehicles in 250-m buffer	28141.99 (SD = 40348.13)	23827.40 (SD = 32150.47)
Heavy-duty vehicles in 250-m buffer	3470.27 (SD = 4492.35)	3284.22 (SD = 4291.09)

*P value <0.05.

usage. BMI was significantly associated with the outcome in both the main and subgroup analyses but with very small effects, possibly due to the number of missing values and the resulting imputations. Moreover, while we used natural language processing to extract EF measurements from the imaging reports and clinical notes, diagnoses in the study data were extracted mainly using structured diagnostic codes. The diagnostic codes used to define HF included chronic and acute HF, as well as both HF with preserved and reduced EF. Therefore, an important limitation of the study is the lack of documentation for NYHA Functional Classification and other biomarkers for HF severity in the EHR.[39] While EF is a measure of cardiac function, it is limited in defining the severity of patient

Table 2. Mixed-effects logistic regression for reduction of EF ($N = 1287$)

	Odds ratio	*P* value	95% confidence interval	
Diabetes	1.046	0.090	0.993	1.101
Medication	1.137	<0.001*	1.076	1.201
Valvular heart disease	0.957	0.083	0.910	1.006
Hypertension	0.996	0.948	0.887	1.119
Smoking	1.054	0.101	0.990	1.122
Male (vs Female)	1.093	<0.001*	1.049	1.139
Race (Base: White)				
Asian	0.915	0.041*	0.840	0.996
Black	0.954	0.292	0.873	1.041
Declined	1.023	0.560	0.948	1.104
Other	0.952	0.088	0.899	1.007
BMI	0.999	0.001*	1.000	1.000
Census-tract poverty rate	1.066	0.440	0.906	1.255
Age	0.997	0.017*	0.995	0.999
Coronary artery disease	1.006	0.885	0.924	1.096
Daily NO_2 concentration ($\mu g/m^3$)	1.071	<0.001*	1.036	1.107

*P value < 0.05.

Table 3. Subgroup analysis of the study cohort whose initial EF was normal: mixed-effects logistic regression for reduction of EF ($N = 1073$)

	Odds ratio	*P* value	95% confidence interval	
Diabetes	1.035	0.215	0.980	1.092
Medication	1.158	<0.001*	1.099	1.221
Valvular heart disease	0.971	0.373	0.909	1.036
Hypertension	0.955	0.435	0.850	1.073
Smoking	1.070	0.062	0.997	1.148
Male (vs female)	1.117	<0.001*	1.065	1.173
Race (Base: White)				
Asian	0.929	0.187	0.833	1.036
Black	0.955	0.291	0.878	1.040
Declined	0.999	0.990	0.915	1.092
Other	0.974	0.407	0.914	1.037
BMI	0.999	0.037*	1.000	1.000
Census-tract poverty rate	1.189	0.053	0.998	1.416
Age	0.997	0.054	0.995	1.000
Coronary artery disease	0.978	0.645	0.891	1.074
Distance (km) to nearest parks	1.166	0.049*	1.001	1.358
Distance (km) to nearest subway stops	0.947	<0.001*	0.927	0.967

*P value < 0.05.

symptoms and not always correlated with physical activity especially in patients with HF with preserved EF. This study conducted a sensitivity analysis using available BNP values near the initial EF measurements to further investigate the patient characteristics in the subgroup. Future studies will need to leverage unstructured data for more accurate data extraction, and also to conduct subgroup analysis on patients with preserved and reduced EF separately. Specifically to this study, in both main and subgroup analyses, we found that decreased age and more medication prescription were significantly associated with an increased odds of reduction in EF. These findings may reflect the characteristics of a patient cohort under treatment in a health system identified from the EHR data in comparison to a general population.

Table 4. BNP values in the patients with normal versus abnormal initial EF measurements

EF category	Mean (SD)	Median
Normal	552.1 (662.2)	352
Abnormal (mild + moderate)	1061.9 (1133.0)	679

This study excluded patients who had recorded address changes during the study period of 2012–2017. As a result, the sample size was a limiting factor in the identification of additional built environment factors that may be significantly associated with the outcome. Despite this effort, there likely still are unrecorded changes in the home locations that were not captured in the data, in addition to exposures prior to 2012 that could have contributed to the study outcome. Future studies may explore large datasets combining EHR data from multiple health systems and insurance claims data to address this challenge. Lastly, our study data did not capture detailed social determinants of health such as individual income, family support, occupation, stress level, and altitude of the apartment buildings that may contribute to the outcome. Future efforts in better tracking social determinants of health in the EHR may alleviate this limitation.

CONCLUSION

Using EHRs linked with address-based metrics on the built environment, we found that air quality, proximity to subway stops, and proximity to parks were associated with a reduction in EF among HF patients in an urban environment. Our findings confirm previous findings on the effects of clean air quality and physical activity for enhanced cardiovascular health. More importantly, findings from this study may help pave the path for promoting future integration of public data sources with EHR data, and more rigorous and precise data collection of patient-level exposure from the built environment in routine patient visits. Clinical decision support may be built within the EHR to provide built environment-related real-time alerts and reminders to care providers for personalized management of HF. Furthermore, collected data may enable larger observational studies on the effects of the built environment on cardiovascular health, thus potentially expediting longitudinal environmental health studies.

FUNDING

This work was supported by the Center for Transportation, Environment, and Community Health New Research Initiatives Fund (79841-10984). This work was partially supported by Weill Cornell Medicine Dean's Diversity and Healthcare Disparities Award, National Library of Medicine (K01LM013257-01), the Clinical and Translational Science Center (UL1 TR000457), Joint Clinical Trials Office, and the University Transportation Research Center September 11th grant (55606-08-28).

AUTHOR CONTRIBUTIONS

Y.Z. designed the overall study in consultation with J.S.A., J.P., and O.G. Y.Z., M.T., E.S. obtained the study data. Y.Z., M.T., E.S., and Y.L. performed data analysis in consultation with A.R.C. S.A.,

A.B., and K.M.A. provided clinical inputs and interpretation. Y.Z. and M.T. wrote the paper with input from all authors.

SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

ACKNOWLEDGEMENTS

We thank Dr. James K. Min, Dr. Andrew Danneberg, Ms. Renee Autumn Ray, and Mr. Brian Stein for their valuable feedback on the study.

CONFLICT OF INTEREST STATEMENT

Y.Z. and J.P. report equity ownership in Iris OB Health, Inc.

DISCLAIMER

The contents of this paper reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The work of coauthors Mohammad Tayarani and H. Oliver Gao in this paper is funded partially by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the US Government assumes no liability for the contents or use thereof.

REFERENCES

1. Galea S, Tracy M. Participation rates in epidemiologic studies. *Ann Epidemiol* 2007; 17 (9): 643–53.
2. Pathak J, Kho AN, Denny JC. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* 2013; 20 (e2): e206–11.
3. Zhang Y, Padman R, Wasserman L, Patel N, Teredesai P, Xie Q. On clinical pathway discovery from electronic health record data. *IEEE Intell Syst* 2015; 30 (1): 70–5.
4. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JP. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (1): 198–208.
5. Perer A, Wang F, Hu J. Mining and exploring care pathways from electronic medical records with visual analytics. *J Biomed Inform* 2015; 56: 369–78.
6. Yao L, Zhang Y, Li Y, Sanseau P, Agarwal P. Electronic health records: implications for drug discovery. *Drug Discov Today* 2011; 16 (13–14): 594–9.
7. Xie S, Greenblatt R, Levy MZ, Himes BE. Enhancing electronic health record data with geospatial information. *AMIA Jt Summits Transl Sci Proc* 2017; 2017: 123–32.
8. Schinasi LH, Auchincloss AH, Forrest CB, Roux AVD. Using electronic health record data for environmental and place based population health research: a systematic review. *Ann Epidemiol* 2018; 28 (7): 493–502.
9. Perdue WC, Stone LA, Gostin LO. The built environment and its relationship to the public's health: The legal framework. *Am J Public Health* 2003; 93 (9): 1390–4.
10. Savarese G, Lund LH. Global public health burden of heart failure. *Card Fail Rev* 2017; 3 (1): 7–11.
11. He J, Ogden LG, Bazzano LA, Vupputuri S, Loria C, Whelton PK. Risk factors for congestive heart failure in US men and women: NHANES I epidemiologic follow-up study. *Arch Intern Med* 2001; 161 (7): 996–1002.
12. Miller KA, Siscovick DS, Sheppard L, *et al.* Long-term exposure to air pollution and incidence of cardiovascular events in women. *N Engl J Med* 2007; 356 (5): 447–58.
13. Cesaroni G, Forastiere F, Stafoggia M, *et al.* Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE Project. *BMJ* 2014; 348: f7412.
14. Pope CA, Burnett RT, Thurston GD, *et al.* Cardiovascular mortality and long-term exposure to particulate air pollution—epidemiological evidence of general pathophysiological pathways of disease. *Circulation* 2004; 109 (1): 71–7.
15. Beelen R, Hoek G, Houthuijs D, *et al.* The joint association of air pollution and noise from road traffic with cardiovascular mortality in a cohort study. *Occup Environ Med* 2008; 66 (4): 243–50.
16. Medina-Ramon M, Goldberg R, Melly S, Mittleman MA, Schwartz J. Residential exposure to traffic-related air pollution and survival after heart failure. *Environ Health Persp* 2008; 116 (4): 481–5.
17. Chum A, O'Campo P. Cross-sectional associations between residential environmental exposures and cardiovascular diseases. *BMC Public Health* 2015; 15 (1): 438.
18. Dominici F, Peng RD, Bell ML, *et al.* Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *JAMA* 2006; 295 (10): 1127–34.
19. Wellenius GA, Bateson TF, Mittleman MA, Schwartz J. Particulate air pollution and the rate of hospitalization for congestive heart failure among Medicare beneficiaries in Pittsburgh, Pennsylvania. *Am J Epidemiol* 2005; 161 (11): 1030–6.
20. Pujades-Rodriguez M, Timmis A, Stogiannis D, *et al.* Socioeconomic deprivation and the incidence of 12 cardiovascular diseases in 1.9 million women and men: implications for risk prediction and prevention. *PLoS One* 2014; 9 (8): e104671.
21. Mannisto T, Mendola P, Grantz KL, *et al.* Acute and recent air pollution exposure and cardiovascular events at labour and delivery. *Heart* 2015; 101 (18): 1491–8.
22. Schuemie MJ, Cepede MS, Suchard MA, *et al.* How confident are we about observational findings in health care: a benchmark study. *Harvard Data Sci Rev* 2020; 2 (1)
23. Hersh WR, Weiner MG, Embi PJ, *et al.* Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med Care* 2013; 51 (8 Suppl 3): S30–7.
24. Lurio J, Morrison FP, Pichardo M, *et al.* Using electronic health record alerts to provide public health situational awareness to clinicians. *J Am Med Inform Assoc* 2010; 17 (2): 217–9.
25. Kruse CS, Stein A, Thomas H, Kaur H. The use of electronic health records to support population health: a systematic review of the literature. *J Med Syst* 2018; 42 (11)
26. Hripcsak G, Duke JD, Shah NH, *et al.* Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
27. Johnson SB, Adekkanattu P, Campion TR Jr, *et al.* From sour grapes to low-hanging fruit: a case study demonstrating a practical strategy for natural language processing portability. *AMIA Jt Summits Transl Sci Proc* 2018; 2017: 104–12.
28. U.S. Department of Commerce. Welcome to Geocoder. https://geocoding.geo.census.gov/ Accessed September 01, 2020.
29. Department of City Planning. NYC Planning. City of New York. 2019. https://www1.nyc.gov/site/planning/data-maps/open-data.page. Accessed September 01, 2020.
30. Vovsha P, Petersen E, Donnelly R. Microsimulation in travel demand modeling: Lessons learned from the New York best practice model. *Transp Res Record* 2002; 1805 (1): 68–77.
31. Karner AA, Eisinger DS, Niemeier DA. Near-roadway air quality: synthesizing the findings from real-world data. *Environ Sci Technol* 2010; 44 (14): 5334–44.
32. Barr J, Cohen JP. The floor area ratio gradient: New York City, 1890–2009. *Reg Sci Urban Econ* 2014; 48: 110–9.
33. Wei YD, Xiao WY, Wen M, Wei R. Walkability, land use and physical activity. *Sustainability Basel* 2016; 8 (1): 65.
34. Frank LD, Sallis JF, Conway TL, Chapman JE, Saelens BE, Bachman W. Many pathways from land use to health: associations between

neighborhood walkability and active transportation, body mass index, and air quality. *J Am Plann Assoc* 2006; 72 (1): 75–87.
35. Muller NZ. Economics boosting GDP growth by accounting for the environment. *Science* 2014; 345 (6199): 873–4.
36. Cantor MN, Chandras R, Pulgarin C. FACETS: using open data to measure community social determinants of health. *J Am Med Inform Assoc* 2018; 25 (4): 419–22.
37. Butler J, Anker SD, Packer M. Redefining heart failure with a reduced ejection fraction. *JAMA* 2019; 322 (18): 1761.
38. Lang RM, Badano LP, Mor-Avi V, *et al.* Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur Heart J Cardiovasc Imaging* 2015; 16 (3): 233–70.
39. New York Heart Association. Criteria Committee. *Nomenclature and Criteria for Diagnosis of Diseases of the Heart and Great Vessels*. Vol. 253: Boston: Little, Brown & Co; 1994.
40. Johnson FL. Pathophysiology and etiology of heart failure. *Cardiol Clin* 2014; 32 (1): 9–19, vii.
41. Ibrahim NE, Burnett JC, Butler J, *et al.* Natriuretic peptides as inclusion criteria in clinical trials: a JACC: heart failure position paper. *JACC Heart Fail* 2020; 8 (5): 347–58.
42. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? *Int J Methods Psychiatr Res* 2011; 20 (1): 40–9.
43. Taylor RS, Sagar VA, Davies EJ, *et al.* Exercise-based rehabilitation for heart failure. *Cochrane Database Syst Rev* 2014; (4): CD003331.
44. O'Connor CM, Whellan DJ, Lee KL, *et al.* Efficacy and safety of exercise training in patients with chronic heart failure: HF-ACTION randomized controlled trial. *JAMA* 2009; 301 (14): 1439–50.
45. Sinharay R, Gong J, Barratt B, *et al.* Respiratory and cardiovascular responses to walking down a traffic-polluted road compared with walking in a traffic-free area in participants aged 60 years and older with chronic lung or heart disease and age-matched healthy controls: a randomised, crossover study. *Lancet* 2018; 391 (10118): 339–49.
46. Barrett MA, Miller D, Frumkin H. Parks and health: aligning incentives to create innovations in chronic disease prevention. *Prev Chronic Dis* 2014; 11: E63.
47. Yuen HK, Jenkins GR. Factors associated with changes in subjective well-being immediately after urban park visit. *Int J Environ Health Res* 2020; 30: 134–145.
48. Torpy JM, Burke AE, Glass RM. JAMA patient page. Acute emotional stress and the heart. *JAMA* 2007; 298 (3): 360.
49. Syed ST, Gerber BS, Sharp LK. Traveling towards disease: transportation barriers to health care access. *J Community Health* 2013; 38 (5): 976–93.