PNAS BRIEF REPORT | ENVIRONMENTAL SCIENCES OPEN ACCESS

Utilizing big data without domain knowledge impacts public health decision-making

Miao Zhang[a], Salman Rahman[a], Vishwali Mhasawade[a], and Rumi Chunara[a,b,1]

Edited by Karen Seto, Yale University, New Haven, CT; received February 4, 2024; accepted July 11, 2024

New data sources and AI methods for extracting information are increasingly abundant and relevant to decision-making across societal applications. A notable example is street view imagery, available in over 100 countries, and purported to inform built environment interventions (e.g., adding sidewalks) for community health outcomes. However, biases can arise when decision-making does not account for data robustness or relies on spurious correlations. To investigate this risk, we analyzed 2.02 million Google Street View (GSV) images alongside health, demographic, and socioeconomic data from New York City. Findings demonstrate robustness challenges; built environment characteristics inferred from GSV labels at the intracity level often do not align with ground truth. Moreover, as average individual-level behavior of physical inactivity significantly mediates the impact of built environment features by census tract, intervention on features measured by GSV would be misestimated without proper model specification and consideration of this mediation mechanism. Using a causal framework accounting for these mediators, we determined that intervening by improving 10% of samples in the two lowest tertiles of physical inactivity would lead to a 4.17 (95% CI 3.84–4.55) or 17.2 (95% CI 14.4–21.3) times greater decrease in the prevalence of obesity or diabetes, respectively, compared to the same proportional intervention on the number of crosswalks by census tract. This study highlights critical issues of robustness and model specification in using emergent data sources, showing the data may not measure what is intended, and ignoring mediators can result in biased intervention effect estimates.

obesity | diabetes | data | Google Street View

Proliferation of digital data, alongside AI and machine learning methods to extract information from them, has potential to inform decision-making which can affect large communities and populations in fields such as public health and urban planning. A growing literature has leveraged object detection via deep learning along with image data such as from Google Street View (GSV) to audit neighborhood properties as well as link them to health outcomes. Environmental and urban development features from GSV data, such as types of vegetation, building structures, and road networks, have been linked to health outcomes and described as useful for informing place-based interventions to improve cardiometabolic diseases, mental distress, and COVID-19 prevalence (1, 2). At the same time, challenges associated with AI-based predictive models have surfaced. These challenges are particularly evident when dealing with nonrepresentative and biased data (3). Additionally, there can be statistical challenges, such as making predictions based on spurious correlations (4). These challenges are amplified when measuring the effects of environmental attributes (such as from GSV data) on health outcomes, where there can be several intermediate factors such as mediators interceding the relationship between these exposures and health outcomes.

In this brief report, we study the association between GSV-derived built environment features and mean obesity and diabetes census tract prevalence in New York City (NYC). We find physical inactivity significantly mediates the relationship and use this causal model to compare the impact of place-based environmental-level interventions (e.g., adding sidewalks) versus behavior-level interventions (e.g., improving physical-activity levels) by census tract. This work illustrates how using emergent data sources in concert with public health domain knowledge is essential for unbiased effect estimation and informing interventions.

Author affiliations: [a]Department of Computer Science and Engineering, Tandon School of Engineering, Brooklyn, NY 11201; and [b]Department of Biostatistics, School of Global Public Health, New York, NY 10003

Author contributions: M.Z., S.R., V.M., and R.C. designed research; M.Z., S.R., and V.M. analyzed data; and M.Z., S.R., V.M., and R.C. wrote the paper.

The authors declare no competing interest.

[1]To whom correspondence may be addressed. Email: rumi.chunara@nyu.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2402387121/-/DCSupplemental.

Published September 17, 2024.

Materials and Methods

Data. Using the Google Application Programming Interface (API), GSV images were collected along all streets in NYC, sampling every 20 m.* At each sampling point images from the four cardinal directions (North, East, South, and West) were collected to form a panorama of the built environment, following previous studies (5). In total, 2.02 million images were collected. Built environment features that may be present in the image–sidewalk and crosswalk–were extracted for all collected images using the Google Vision API.† Each image is labeled with sidewalk or crosswalk presence if the API returns a probability score bigger than 0.8 for the label "sidewalk" or the label "zebra crossing" (label names defined by the API); otherwise no label is applied. The GSV-estimated feature prevalence is computed by taking the ratio of labeled sampling points to all sampling points within each census tract. Sidewalk ground truth data distribution is obtained from NYC OpenData from 2023,‡ in which each sidewalk is recorded as a polygon. We compute the prevalence by dividing the area of all sidewalk polygons of a census tract by the area of the census tract, referred to as ground truth sidewalk prevalence. Feature representativenss is computed via correlation analysis (*SI Appendix*).

Health outcome and health-related behavior data were obtained from the Centers for Disease Control and Prevention PLACES initiative (6) at the census tract level ($n = 1{,}970$). We selected the health outcome measures of "Obesity among adults aged $\geq$ 18 y" (denoted as obesity) and "Diabetes among adults aged $\geq$ 18 y" (denoted as diabetes). The health-related behavior "No leisure-time physical activity among adults aged $\geq$ 18 y" (denoted as physical inactivity) is tested as mediator. The same demographic and socioeconomic controlling factors from previous work (7) were used, including socioeconomic status; percentage female, white, and Hispanic; percentage <18 and $\geq$65, and median age year.

The choice of built environment features and health outcomes is motivated by the relevance of sidewalks and crosswalks to obesity and diabetes (8), as in previous analyses on GSV data (7, 9). Physical inactivity is used as the mediator, as it reflects lifestyle habits positively related to risk of the studied health outcomes (10, 11). Moreover, physical inactivity has been shown to mediate the effect of built environment features like walkability (12), land-use mix (13), and residential blue space (14), on health outcomes such as obesity.

Model and Counterfactual Outcome. A mediation model framework was used (Fig. 1*A*, *SI Appendix*). To investigate the effect of disregarding mediation of environmental attributes measured via GSV on decision-making, we investigated the change in outcome (mean value of obesity and diabetes) via intervention at the i) environment level, and ii) physical activity level. For the built environment intervention, we split samples into three tertiles: high, moderate, and low based on the distribution of X, and define 1 unit of intervention as setting 10% of samples in low and moderate tertile to have the same mean value of X as the high tertile. The intervention on individual-level behavior is performed with the same procedure on the M distribution. The counterfactual Y after intervention is computed using coefficients of the established mediation model (*SI Appendix*).

Results

GSV-Derived Features Association with Health Outcomes. The total effect of the crosswalk feature, as measured by GSV, is negative, indicating that a higher crosswalk density is associated with lower disease prevalence. Further, the effect on health outcomes is significant, with a bigger effect on obesity ($c = -1.11$) than diabetes ($c = -0.153$). These findings are consistent with previous studies based on GSV-estimated crosswalk feature (2, 7, 9). However, no significant associations were found between GSV-estimated sidewalk feature and health outcomes (Table 1), in contrast to a previous study which analyzed the association nation-wide (15).

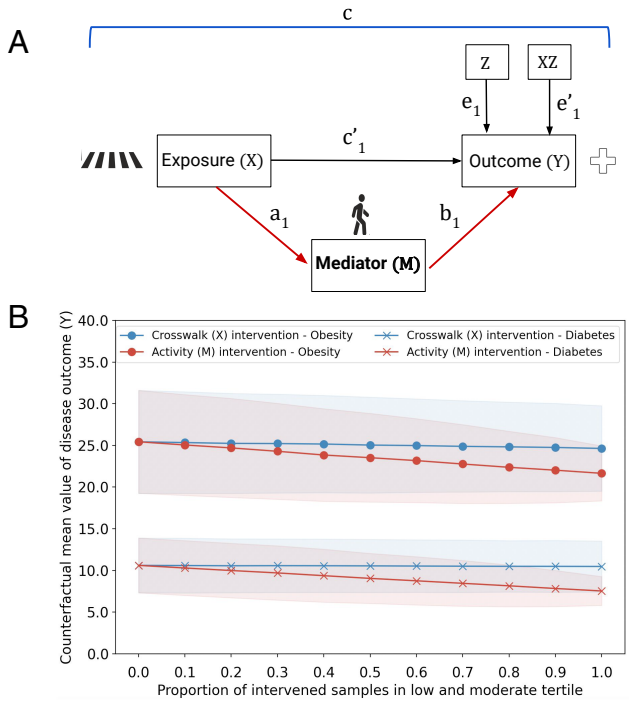

Fig. 1. (*A*) Path diagram of the mediation model. Total effect, c, of crosswalk exposure (X) on the outcome (Y), is decomposed into an indirect effect via physical inactivity mediator (M), quantified by a1b1, and direct effect quantified by c'1. (*B*) Performing health outcome improvement intervention: X and M intervenable variables are grouped into three tertiles; tracts in the low and moderate tertiles are altered to have the same mean as that in the high tertile. The counterfactual Y changes after intervening on X or M, and the outcome mean and $\pm$ SD σ is plotted: 1 unit intervention on M (setting 10% of samples in the two lowest tertiles to have the same mean value of the highest) would result in 4.17 (95% CI: 3.84, 4.55) times bigger decrease on obesity and 17.2 (95% CI: 14.4, 21.3) times bigger decrease on diabetes, than the same unit intervention on X. Accounting for the mediation mechanism allows comparative assessment of intervention efficacy, showing a proportional intervention on physical inactivity has a larger effect on the outcome than the same proportional intervention on the built environment exposure.

Physical Inactivity Is a Significant Mediator. All mediation effect ($a_1 b_1$) directions are negative, operating through a negative exposure effect on the mediator (a_1) and positive effect of the mediator on the outcome (b_1) (Table 1). The total effect of GSV-measured crosswalk prevalence is significantly mediated by physical inactivity prevalence ($a_1 b_1 = -2.46/-1.96$, for obesity/diabetes), and physical inactivity entirely mediates the total effect of sidewalk prevalence on health outcomes ($a_1 b_1 = -0.996/-0.721$), as neither simple association between sidewalk prevalence and obesity/diabetes are significant (Table 1). That is, decreased obesity or diabetes prevalence linked to increased crosswalk or sidewalk prevalence by census tract, can be largely accounted for by whether individuals in that census tract have increased physical activity.

Intervening on Physical Inactivity Versus Street View Features. Analysis shows that a 1 unit decrease of physical inactivity would result in a 4.17 (95% CI: 3.84, 4.55) times bigger decrease of obesity prevalence and 17.2 (95% CI: 14.4, 21.3) times bigger decrease of diabetes prevalence, compared to the same intervention on crosswalk prevalence (Fig. 1*B*).

Street View Features May Not Represent the Built Environment. Pearson correlation coefficient between GSV-estimated and ground truth sidewalk prevalence was 0.214 ($P < 0.001$)

*https://developers.google.com/maps/documentation/streetview/overview.

†https://cloud.google.com/vision/docs/labels.

‡https://data.cityofnewyork.us/City-Government/Sidewalk/vfx9-tbb6.

Table 1. Effect sizes and 95% CIs (in brackets) for the model that tests the total effect of built environment exposure (*X*): Crosswalk and Sidewalk on health outcome (*Y*): Obesity and Diabetes, and the models that test the mediation effect of inactivity (*M*) in the relationship of exposure with outcome

	Total effect (c)	Mediation effect ($a_1 b_1$)	Effect on mediator (a_1)	Mediator effect on outcome (b_1)
Health outcome: Obesity				
Crosswalk	−1.11***	−2.46***	−0.682***	3.60***
	[−1.36, −0.860]	[−3.58, −1.50]	[−0.896, −0.468]	[3.21, 4.00]
Sidewalk	0.00125	−0.996**	−0.253**	3.93***
	[−0.187, 0.189]	[−1.82, −0.305]	[−0.420, −0.0858]	[3.55, 4.33]
Health outcome: Diabetes				
Crosswalk	−0.153***	−1.96***	−0.682***	2.88***
	[−0.243, −0.0626]	[−2.67, −1.30]	[−0.896, −0.468]	[2.77, 2.98]
Sidewalk	0.0246	−0.721**	−0.253**	2.85***
	[−0.0489, 0.0981]	[−0.236, −1.24]	[−0.420, −0.0858]	[2.75, 2.95]

***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. Note that Effect on mediator (a_1) is independent of health outcome.

at the city level. At borough level, the correlation is significant ($P < 0.05$) with coefficient 0.361 for Bronx, 0.386 for Manhattan, 0.327 for Queens, and 0.331 for Staten Island, but not significant for Brooklyn. Qualitative examination shows GSV may falsely report sidewalks near highways and bridges, or absence in places with sidewalk obstructions (Fig. 2).

Discussion

Growing amounts of digital data can be useful to inform decision-making, but our key finding is that simply using associations as suggested in previous work, can misappropriate the utility of the information. First, we showed that mean physical inactivity by census tract significantly mediates the impact of built environment features measured through GSV, on prevalence of obesity and diabetes. We then showed that accounting for mediators is important for accurate targeting and efficacy estimation of interventions.

Our study makes several important advances. Importantly, we examine GSV compared to ground truth data at city level, opposed to existing studies which compare areas by qualitative review. In doing so, we show that at a granular level, built environment features based on labels from GSV may not match with ground truth. We also utilize a causal framework to show how informing interventions based on the data must take such mechanisms into account; especially in light of common underfunding to public health. The mediation effect was also significant with ground truth data. However, caution is advised when interpreting the effect magnitudes reported for GSV due to data robustness issues described herein.

In conclusion, this study highlights the potential of digital data sources like GSV in enhancing public health research, while also pointing out the need for careful consideration of data limitations and the complex dynamics between the built environment, individual behavior, and health outcomes. Future research should focus on addressing these challenges and exploring innovative strategies to leverage digital data for more effective public health interventions.

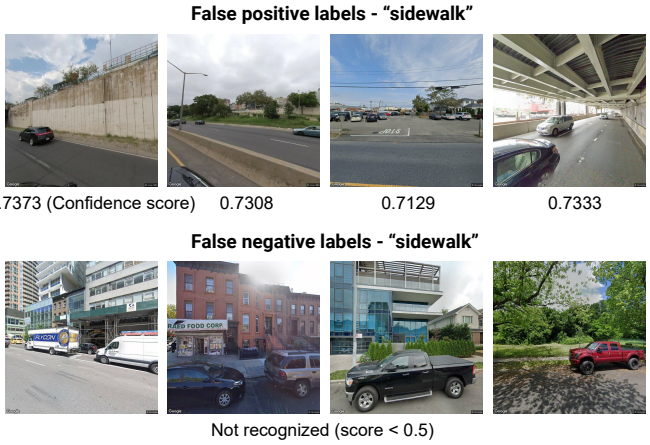

Fig. 2. Common false positive and negative errors for GSV labels for sidewalk. *Top*: GSV incorrectly returns high confidence of a sidewalk, as side lanes of highways, bridges, or parking lots have similar shape to and are easily confused with sidewalks. *Bottom*: GSV labels have incorrect low confidence recognizing sidewalk because of obstruction from vehicles, constructions, or tree shade. Data source: GSV images.

Data, Materials, and Software Availability. Data cannot be shared [work using Google Street View data is considered fair use under their policy (https://about.google/brand-resource-center/products-and-services/geo-guidelines/), and therefore we cannot share the data used per Google's terms].

ACKNOWLEDGMENTS. NSF Grant 1845487.

1. Q. C. Nguyen *et al.*, Using 164 million Google Street View images to derive built environment predictors of COVID-19 cases. *Int. J. Environ. Res. Public Health* **17**, 6359 (2020).
2. J. M. Keralis *et al.*, Health and the built environment in united states cities: Measuring associations using Google Street View-derived indicators of the built environment. *BMC Public Health* **20**, 1–10 (2020).
3. J. Buolamwini, T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification" in *Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
4. A. J. DeGrave, J. D. Janizek, S. I. Lee, AI for radiographic COVID-19 detection selects shortcuts over signal. *Nat. Mach. Intell.* **3**, 610–619 (2021).
5. J. H. Kim, S. Lee, J. R. Hipp, D. Ki, Decoding urban landscapes: Google Street View and measurement sensitivity. *Comput. Environ. Urban Syst.* **88**, 101626 (2021).
6. K. J. Greenlund *et al.*, PLACES: Local data for better health. *Prev. Chronic Dis.* **19**, E31 (2022).
7. Q. C. Nguyen *et al.*, Leveraging 31 million Google Street View images to characterize built environments and examine county health outcomes. *Public Health Rep.* **136**, 201–211 (2021).
8. J. Wei *et al.*, Neighborhood sidewalk access and childhood obesity. *Obesity Rev.* **22**, e13057 (2021).
9. Q. C. Nguyen *et al.*, Google Street View images as predictors of patient health outcomes, 2017–2019. *Big Data Cognit. Comput.* **6**, 15 (2022).
10. J. M. Jakicic, K. K. Davis, Obesity and physical activity. *Psych. Clin.* **34**, 829–840 (2011).
11. D. C. Lee *et al.*, Leisure-time running reduces all-cause and cardiovascular mortality risk. *J. Am. Coll. Cardiol.* **64**, 472–481 (2014).
12. J. Van Cauwenberg, V. Van Holle, I. De Bourdeaudhuij, D. Van Dyck, B. Deforche, Neighborhood walkability and health outcomes among older adults: The mediating role of physical activity. *Health Place* **37**, 16–25 (2016).
13. Y. Xiao, S. Chen, S. Miao, Y. Yu, Exploring the mediating effect of physical activities on built environment and obesity for elderly people: Evidence from Shanghai, China. *Front. Public Health* **10**, 853292 (2022).
14. T. P. Pasanen, M. P. White, B. W. Wheeler, J. K. Garrett, L. R. Elliott, Neighbourhood blue space, health and wellbeing: The mediating role of different types of physical activity. *Environ. Int.* **131**, 105016 (2019).
15. X. Yue *et al.*, Using convolutional neural networks to derive neighborhood built environments from Google Street View images and examine their associations with health outcomes. *Int. J. Environ. Res. Public Health* **19**, 12095 (2022).