BMJ Public Health

Machine learning to evaluate the relationship between social determinants and diabetes prevalence in New York City

Darren Tanner ,[1] Yongkang Zhang,[2] Ji Eun Chang,[3] Peter Speyer,[4] Elizabeth Adamson,[4] Ann Aerts,[4] Juan M Lavista Ferres,[1] William B Weeks[1]

To cite: Tanner D, Zhang Y, Chang JE, *et al*. Machine learning to evaluate the relationship between social determinants and diabetes prevalence in New York City. *BMJ Public Health* 2024;**2**:e001394. doi:10.1136/bmjph-2024-001394

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/bmjph-2024-001394).

Received 29 April 2024
Accepted 9 September 2024

[1]AI for Good Research Lab, Microsoft Corp, Redmond, Washington, USA
[2]Department of Population Health Sciences, Weill Cornell Medicine, New York, New York, USA
[3]New York University School of Global Public Health, New York, New York, USA
[4]Novartis Foundation, Basel, Switzerland

Correspondence to
Dr Darren Tanner;
darren.tanner@microsoft.com

ABSTRACT

Introduction Diabetes is a leading contributor to cardiovascular disease and mortality; social determinants of health (SDOH) are associated with disparities in diabetes risk. Quantifying the cumulative impact of SDOH and identifying the SDOH most associated with diabetes prevalence at the neighbourhood level can help policy-makers design and target local interventions to mitigate these disparities. Machine learning (ML) methods can provide novel insights and help inform public health intervention strategies in a place-based manner.
Methods In a cross-sectional study, we used gradient boosting ML models to estimate the cumulative contribution of a set of SDOH variables to diabetes prevalence (%) at the census tract level within New York City (NYC); Shapley Additive Explanations were used to assess the magnitude and shape of relationships between our SDOH variables and model-predicted NYC diabetes prevalence. SDOH measures included socioeconomic position, educational attainment, food access, air quality, neighbourhood environment, housing conditions and insurance coverage.
Results Across 2096 NYC census tracts (population 8 170 505), mean diabetes prevalence was 11.5% (SD 3.7%; range 1.9%–42.8%). A set of 16 SDOH variables representing a framework of 16 distinct SDOH concepts accounted for 67% of the between-tract variance in model-derived NYC diabetes prevalence estimates (95% CI 66% to 68%); a set of 81 variables representing these 16 concepts accounted for 80% of variance (95% CI 78% to 81%). Models showed excellent across-location generalisation. The most important variables driving model predictions within NYC were measures of low educational attainment and poverty.
Conclusions SDOH accounted for a substantial proportion of neighbourhood-level variation in diabetes prevalence within NYC, independent of the demographics and health behaviours associated with those SDOH. Our place-based findings suggest that, within NYC, where approximately one million residents have diabetes and there are legislative requirements to reduce the impacts from diabetes, policies reducing socioeconomic and educational inequality could have the greatest potential to equitably achieve this.

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Social determinants of health (SDOH) are associated with chronic diseases, including diabetes. Although numerous individual SDOH have been found to be associated with disproportionate diabetes prevalence, there is little evidence regarding local cumulative impact and relative contribution of these SDOH when considering a constellation of SDOH simultaneously. Additionally, most prior work does not consider potential interactions and non-linear relationships between SDOH and diabetes.

WHAT THIS STUDY ADDS

⇒ We demonstrate that non-linear, interactive machine learning models show improved accuracy and estimate a larger cumulative association between SDOH and between-neighbourhood disparities in diabetes prevalence within New York City (NYC) than did traditional statistical methods. Additionally, our work demonstrates how policy-makers can harness machine learning in a place-based way to assist in designing local initiatives and interventions in a targeted, data-driven way.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ NYC has recently enacted legislation requiring the Department of Health and Mental Hygiene to implement initiatives that reduce the impacts of diabetes and extend life expectancy; our NYC place-based findings suggest that initiatives focused on reducing poverty and increasing educational attainment might have the largest impact in reducing diabetes prevalence there. Other jurisdictions could adopt these place-based methods to identify leading SDOH to target for public health interventions.

INTRODUCTION

Social determinants of health (SDOH) are the non-medical factors that influence health outcomes; SDOH encompass the conditions in which people are born, grow, work, live and age, as well as the wider set of forces and systems shaping the conditions of daily life.[1] SDOH can shape individuals' health behaviours, which can then influence health outcomes.[2] In the USA, SDOH account for up to 50% of health outcomes and have been established as drivers of outcome inequities across numerous health conditions.[3–5] Variation in SDOH across neighbourhoods has increased in recent decades as a result of the widening distribution of economic prosperity among US communities.[6] These increasing inequalities have expanded interest in developing policies and resources that support both 'places' and 'people', particularly in areas disproportionately affected by disadvantaged social conditions.[7 8] Such policies recognise that socioeconomic conditions are significant determinants of health and that SDOH-based interventions may improve population health.[9 10]

Cardiovascular disease (CVD) is the leading cause of death in the USA and in the world.[11 12] The American Heart Association considers diabetes to be one of the seven major controllable risk factors for CVD,[13] and there is evidence that numerous dimensions of SDOH are associated with diabetes risk.[14] Understanding and addressing the impact of SDOH has been a research priority due to the high and increasing prevalence of this disease.[15–17] Based on four leading theoretical frameworks of SDOH, a 2021 systematic review[14] found a body of research showing strong associations between various SDOH and incidence, prevalence, and outcomes of diabetes. These include socioeconomic status (eg, education and income), neighbourhood and physical environment (eg, housing and toxic exposures), food environment (eg, food insecurity and food access), healthcare (eg, access and quality) and social context (eg, social cohesion and social capital). These findings highlight the importance of addressing disadvantages in SDOH to improve population health, including ameliorating disparities in diabetes risk and prevalence.

Nonetheless, the potential to harness SDOH to reduce diabetes and CVD remains largely unrealised, as few interventions have addressed economic stability, education access and quality, or community context and social risk.[14 18] Additionally, most studies on the relationship between SDOH and diabetes have focused on identifying the associations between diabetes and a single or a limited set of SDOH exposures at a time, without considering the multidimensionality of SDOH and the complex interactions among SDOH.[14] Further, the association between SDOH and health outcomes may differ by geographical region[19]; area-specific evidence is therefore needed to target local interventions aimed at mitigating the impact of SDOH disadvantages on diabetes risk and outcomes. Data-driven approaches that account for these complex relationships among multiple SDOH and that can also target area-specific needs are therefore needed to identify those SDOH where policy interventions might have the largest and most sustained impact on health equity.

Motivated by three primary reasons, we sought to determine the potential combined impact of a comprehensive set of SDOH on across-census tract variation in diabetes prevalence in New York City (NYC), as well as identify the most important SDOH measures that explain this variation. First, with a population of over 8 million, NYC is the largest city in the USA, where nearly 1 million residents have diabetes.[20] Targeted interventions there could therefore benefit a substantial number of people. Second, NYC's Council recently passed laws requiring the Department of Health and Mental Hygiene (DOHMH) to implement data-driven citywide initiatives to reduce the impacts of diabetes and to extend life expectancy.[21–23] Third, the NYC DOHMH is a stakeholder in the Novartis Foundation's AI4HealthyCities initiative,[24] a public–private partnership that seeks to use advanced analytics to decipher drivers of cardiovascular risk and implement targeted interventions to reduce CVD. Insights about the overall impact of SDOH on diabetes prevalence and the strongest SDOH drivers within NYC can be used by policy-makers within these initiatives, potentially leading to tangible public health benefits.

To address these goals, we used machine learning (ML). In contrast to the most commonly used epidemiological statistical approaches, ML can model complex, non-linear relationships and higher-order interactions between SDOH, providing good approximations to the dynamics of real-world systems[25]; given how SDOH interact in real-world contexts, ML can provide an excellent estimate of their cumulative contribution to diabetes. Additionally, ML interpretation algorithms provide insight into what drives model predictions while maintaining a model's non-linear, interactive properties[26]; this can help policy-makers understand which SDOH are most associated with diabetes prevalence within NYC when considering all variables in the model, along with their interactions, simultaneously.

METHODS

Data sources and study population

Data for this cross-sectional study were drawn from publicly available sources at the census tract level using 2010 tract boundaries. The outcome measure, tract-level diabetes prevalence (%), was obtained from the US Centers for Disease Control and Prevention (CDC) PLACES project[27] 2020 data release,[28] reflecting data from the 2018 CDC Behavioral Risk Factor Surveillance System survey (BRFSS).[29] In order to ensure within-tract sampling adequacy, we subset to tracts with populations of at least 500. We further subset the tracts outside NYC to only those that were included in the CDC's 500 Cities project's 2018 data release,[30 31] so that non-NYC tracts used for model training reflect areas with some urbanisation.

Tract-level SDOH predictor variables were compiled from a set of publicly available sources: the US Agency for Healthcare Research and Quality database,[32] the CDC Agency for Toxic Substances Disease Registry 2022 Environmental Justice Index[33] and the US Department of Agriculture Food Access Research Atlas 2019 release.[34] We follow reporting guidelines for artificial intelligence in medical and scholarly research.[25 35]

SDOH measures

Diabetes is associated with multiple neighbourhood, health system, patient and other characteristics, with complex relationships between these factors.[14] We adapted a conceptual framework developed by Brown *et al*[36] to guide our analysis. The framework includes both direct and indirect associations between SDOH and diabetes through three important pathways (online supplemental figure 1). Specifically, SDOH will influence diabetes through health knowledge and behaviours, access to care and behavioural health, and through psychological or other medical conditions. These factors are considered potential mediators between SDOH and diabetes. Demographics (eg, age and sex) are associated with diabetes through direct and multiple indirect pathways, including SDOH. In this study, we examined total associations (including both direct and indirect) between SDOH and diabetes; we therefore excluded mediators within this framework from our analysis. As our goal was to identify SDOH that were potentially modifiable through policy intervention, we also excluded demographic variables.

Guided by the conceptual framework and relevant literature,[14 37–39] we used the following steps to select variables for inclusion in our analysis. First, we identified 17 SDOH concepts which have been shown to relate to diabetes. We required that each concept has a conceptual justification and empirical support from previous work, and that it is not considered a potential mediator between SDOH and diabetes within our adopted framework. Second, we mapped SDOH measures from the public data sources described above to each of the 17 concepts. As no relevant national tract-level crime data were readily publicly available, we excluded this concept, leaving the remaining 16 concepts for further analysis (see online supplemental table 1). There were 81 SDOH measures, with 1–14 measures per concept. Finally, to improve the interpretability of the final models, we reduced the set of variables to one per SDOH concept through a set of heuristic and statistical assessments to identify variables that captured large segments of the population, demonstrated substantial variance across census tracts, were exemplary of their respective pillar, and showed strong statistical associations with diabetes prevalence when considering only census tracts outside of NYC (the training data; see the 'Model development and evaluation' section). Details on this process are provided in online supplemental methods. The final set of 16 concepts, conceptual justifications, variables and variable definitions are provided in online supplemental table 1; see online supplemental table 2 for the set of candidate variables prior to the final exclusions.

Outcome

The outcome measure was census tract-level diabetes prevalence (%), obtained from the CDC PLACES 2020 data files, described above. This reflects the percentage of respondents to the 2018 BRFSS who reported having ever been told by a doctor, nurse, or other health professional that they have diabetes other than diabetes during pregnancy, excluding those who refused to answer, had a missing answer or answered 'don't know/not sure'.[40]

Model development and evaluation

We used ML to examine the association between SDOH and diabetes prevalence and to identify the SDOH most influential in accounting for between-tract variability in diabetes prevalence. We followed standard ML practice by excluding the data used for model evaluation from all stages of variable selection, model design, and model training. We therefore split the data into model training and test sets based on location, with data from all non-NYC tracts, after subsetting as described above, used for variable selection and model training, and data from NYC tracts after subsetting used for model testing/evaluation. Splitting by location mitigates risks of bias where variable selection and model parameters may be influenced by data biases within NYC itself or by overestimating model performance due to spatial autocorrelation. Performance metrics are therefore conservative, as they estimate how our models generalise to a new location.

We used gradient-boosted regression trees (GBRTs), implemented with the XGBoost package.[41] GBRTs are ensembles of regression trees, where successive trees iteratively reduce model errors from the previous trees and include both non-linearities and higher-order interactions in the predictor–outcome relationship. As XGBoost can handle missing data natively, no imputation was done; GBRTs are invariant to data scale and make no distributional assumptions, so no transformations were done.

Two models were trained, one using the narrow set of 16 variables and one using the full set of 81 candidate variables after initial screening. The goal of the first model, which is our primary focus here, was to balance model predictive accuracy and interpretability by including only one variable per SDOH concept; the goal of the second was to assess the broad predictive power of SDOH for diabetes prevalence using our SDOH framework and publicly available data. The training objective was minimising root mean squared error. See online supplemental methods and online supplemental table 3 for hyperparameter optimisation information. Performance was assessed by predicting diabetes prevalence in NYC (the test dataset) using the fitted models. Objective performance metrics were the proportion of variance explained (R^2) and mean absolute error (MAE) for diabetes prevalence over the test dataset; 95% CIs were obtained via bootstrapping (see online supplemental methods).

Interpretation of variable importance and the shape of the relationship between the predictor variables and model predictions used the Shapley Additive Explanations (SHAP) algorithm.[42] SHAP provides a unique value per predictor variable for each observation (census tract), giving the amount that observation's value for that variable pushed the model's prediction away from the dataset mean, while considering all other variables and interactions in the model. SHAP values above 0 indicate that the observation's value for that input variable moved the observation's prediction above the dataset mean; SHAP values below 0 indicate that the variable's value pushed the observation's prediction below the mean. As our outcome measure is diabetes prevalence in percentage points, SHAP values are similarly on the scale of percentage points of diabetes prevalence. Variables with larger mean absolute SHAP values contribute more to model predictions. Normalised mean absolute SHAP values sum to 1 across all variables in the model and reflect the proportion of variance each variable contributes to model predictions by pushing individual predictions away from the mean.

As a comparison baseline for the GBRT models, we also fit traditional ordinary least squares (OLS) regression models both for the 16 variable and 81 variable sets. These were fit using the same approach as above by estimating model parameters from the non-NYC dataset and predicting values for the NYC dataset. As OLS cannot handle missing values, we fit models for each variable set using two approaches: by dropping observations with any missing values among the predictor variables and by interpolating missing values. Interpolation was done using the IterativeImputer class from the scikit-learn package (V.1.5.0) for Python, with random forest regressors as the base estimators. The imputation estimators were fit on the non-NYC training data and used to impute values for both the non-NYC and NYC datasets. 95% CIs for the R^2 and MAE metrics from the OLS models for each missing value approach (dropping or imputing) and each variable set (16 or 81) were computed using the same bootstrap method as above.

All modelling was carried out in Python (V.3.10.13). XGBoost modelling, including SHAP value calculation, was carried out using the XGBoost package (V.2.0.3); OLS models were fit using the statsmodels package (V.0.14.2). Statistical significance was considered at p<0.05, two sided.

Patient and public involvement

As a partner in the Novartis Foundation's AI4HealthyCities initiative, in September 2023, the NYC DOHMH provided feedback to the authors of this study suggesting diabetes prevalence as the preferred outcome variable. This suggestion was made to provide timely, actionable information on relevant SDOH and their relationship with diabetes to support their ongoing legally mandated initiatives to track and reduce the impacts of diabetes.

RESULTS

After subsetting census tracts as described above, there were 25 338 non-NYC tracts used for model training (population 106 824 787; mean diabetes prevalence, 11.2% (SD 4.5%; range 1.0–35.7%)) and 2096 NYC tracts used for model evaluation and variable importance estimation (population 8 170 505; mean diabetes prevalence, 11.5% (SD 3.7%; range 1.9%–42.8%)). We focus here primarily on results from the narrow 16-variable model. Distributions for diabetes prevalence and each of the 16 narrow variables by quintile of diabetes prevalence for NYC are shown in table 1. NYC census tracts with higher diabetes prevalence showed SDOH disadvantages compared with those with lower prevalence. For example, the mean unemployment rate among census tracts in the highest diabetes prevalence quintile was over twice that in the lowest quintile. Distribution information for all variables is shown in online supplemental table 4 and 5 for the non-NYC (training) and NYC data, respectively. Density plots of the outcome and 16 narrow predictor variables for the training and test distributions are depicted in online supplemental figure 2 and 3.

When evaluated against the observed data, our narrow GBRT model, trained on non-NYC tracts, accounted for 67% of the between-tract variance in NYC diabetes prevalence (R^2=0.67, 95% CI 0.66 to 0.68); the average difference between predicted and observed diabetes prevalence was ±1.57 percentage points (MAE=1.57, 95% CI 1.56 to 1.61). The all-variables GBRT model showed statistically significant improvements over the narrow model (p<0.05; figure 1). We found good correspondence between the observed and GBRT-predicted values' spatial distributions (figure 2).

Our ML-based GBRT models performed better than the baseline models fit using traditional OLS regression. The narrow 16 variable OLS model with missing values dropped accounted for 63% of variance in the data and had an average error of ±1.73 percentage points (R^2=0.63, 95% CI 0.63 to 0.64; MAE=1.73, 95% CI 1.72 to 1.75); the model with missing values imputed accounted for 63% of variance and had an average ±1.74 percentage point error (R^2=0.63, 95% CI 0.63 to 0.64; MAE=1.74, 95% CI 1.72 to 1.76). The OLS model using the full 81 variable set and with missing values dropped accounted for 64% of variance in the NYC data and had an average error of ±1.45 percentage points (R^2=0.64, 95% CI 0.61 to 0.67; MAE=1.45, 95% CI 1.40 to 1.53); the model with missing values imputed accounted for 70% of variance with an average error of ±1.58 percentage points (R^2=0.70, 95% CI 0.67 to 0.72; MAE=1.58, 95% CI 1.51 to 1.66). Thus, our non-linear, interactive ML approach showed lower error than traditional linear models in estimating tract-level diabetes prevalence in NYC when comparing the respective 16 and 81 variable models, and they also suggested a larger cumulative association between SDOH and diabetes prevalence than did traditional linear models.

SHAP variable importance metrics for the narrow model (table 2) show that, of the 16 predictor variables, the most influential were low educational attainment, enrolment in the supplemental nutritional assistance programme (SNAP), household broadband and proportion of older adults living alone. These four variables each contributed at least 10% and combined accounted for 62% of the variation in model predictions; the remaining 12 narrow variables each contributed less than 6% to

Table 1 Diabetes prevalence and SDOH variable characteristics for New York City by diabetes prevalence quintile

		Mean (SD)					
			Diabetes prevalence quintile				
Variable	SDOH concept	Total	Quintile 1	Quintile 2	Quintile 3	Quintile 4	Quintile 5
Diabetes prevalence (%)	N/A (outcome)	11.49 (3.71)	6.39 (1.58)	9.83 (0.60)	11.54 (0.46)	13.30 (0.56)	16.68 (2.49)
Mean days >EPA* PM2.5† standard	Air quality	9.29 (0.22)	9.46 (0.15)	9.29 (0.17)	9.25 (0.21)	9.19 (0.26)	9.25 (0.22)
% commute <15 min	Commute	9.27 (6.61)	12.05 (9.04)	9.44 (5.50)	8.88 (6.57)	7.21 (4.63)	8.69 (5.44)
% high school education or less	Education	43.97 (17.66)	20.65 (13.66)	40.58 (11.11)	48.04 (11.86)	52.12 (10.50)	59.67 (10.23)
% unemployment	Employment	7.14 (4.31)	4.59 (2.48)	5.68 (2.65)	6.45 (2.85)	8.18 (4.07)	10.96 (5.55)
% limited English	English proficiency	15.02 (13.21)	7.26 (6.40)	14.34 (10.99)	17.95 (12.60)	14.69 (13.58)	21.39 (16.28)
% essential labour force	Essential occupations	18.15 (7.48)	11.32 (5.78)	19.63 (6.92)	20.71 (7.02)	20.73 (6.96)	18.60 (6.20)
Low food access @ half mile	Food access	0.08 (0.28)	0.05 (0.22)	0.16 (0.37)	0.09 (0.29)	0.07 (0.25)	0.04 (0.20)
% uninsured	Health insurance	8.45 (5.38)	5.16 (3.66)	7.68 (5.10)	9.58 (5.38)	10.02 (4.97)	9.99 (5.89)
% housing w/o complete kitchen	Housing conditions	1.70 (2.25)	1.54 (2.09)	1.54 (1.83)	1.82 (2.32)	1.91 (2.29)	1.72 (2.64)
% SNAP‡ benefits	Income	19.79 (15.74)	6.81 (7.14)	12.87 (8.89)	18.83 (10.90)	22.71 (11.67)	38.65 (16.77)
Gini index	Income inequality	0.46 (0.07)	0.48 (0.07)	0.45 (0.06)	0.45 (0.06)	0.45 (0.06)	0.49 (0.07)
% households w/ broadband	Internet/mobile access	79.22 (10.61)	87.03 (8.28)	81.76 (8.11)	79.47 (10.61)	77.76 (7.73)	69.61 (9.95)
% over 65 living alone	Living arrangement	10.37 (5.99)	10.99 (6.06)	10.54 (5.33)	9.32 (4.88)	8.86 (4.92)	12.13 (7.76)
% rent >30% household income	Rent/mortgage cost	54.82 (13.21)	44.25 (11.20)	52.49 (12.77)	57.57 (10.82)	59.54 (12.64)	60.80 (10.97)
% single parent families	Single-parent households	33.59 (22.83)	17.35 (13.77)	24.25 (15.48)	30.24 (19.27)	40.65 (19.82)	56.53 (22.21)
EPA National walkability index	Walkability	14.02 (1.84)	14.30 (1.94)	14.19 (1.72)	14.06 (1.83)	14.11 (1.71)	13.44 (1.90)

*US Environmental Protection Agency.
†Fine particulate matter ≤2.5 μm in diameter.
‡Supplemental Nutrition Assistance Programme.
N/A, not applicable; SDOH, social determinants of health.

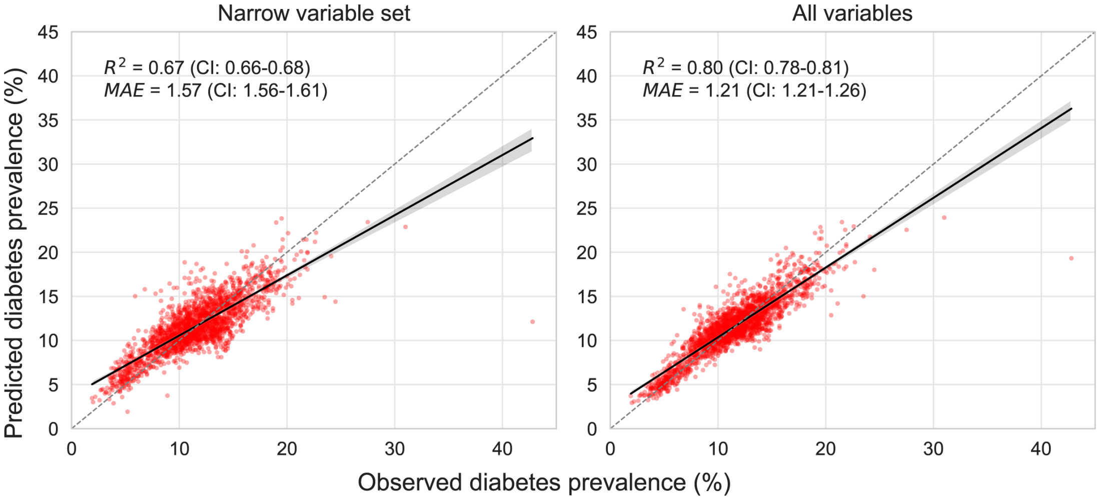

Figure 1 Observed and model-predicted diabetes prevalence within New York City (NYC). Results for both narrow model with 16 SDOH predictors (left) and all-variables model with 81 SDOH predictors (right). Each point represents one NYC census tract. Solid lines show least squares regression fits for predicted-observed relationships; grey bands reflect 95% CIs. Dashed lines show optimal hypothetical fit where predicted and observed values are equal. MAE, mean absolute error; SDOH, social determinants of health.

prediction variation. The top three variables (education, SNAP and broadband) were all strongly intercorrelated (Spearman |ρ|>0.75 in the training set), suggesting they reflect a shared underlying construct of socioeconomic deprivation. See online supplemental table 6 for SHAP values from the all-variables model.

Scatterplots showing the relationship between SHAP values and the top eight most influential variables from the narrow model (figure 3) show that predictions of increased diabetes prevalence were associated with higher rates of poor educational attainment, enrolment in SNAP, proportion of older adults living alone, single parent families and housing units without complete kitchen facilities. Lower rates of household broadband, shorter commute times and less walkable neighbourhoods were also associated with higher model-predicted diabetes prevalence. Additionally, some of these relationships showed clear non-linearities, for example, the increase in predicted diabetes prevalence for rates of high school education or less flattened above approximately 50%, the increase showed an inflexion at around 20% rates of SNAP participation and the decrease associated with higher household broadband rates flattened above approximately 75%.

DISCUSSION

Using ML, we found that a small set of SDOH measures can account for a substantial proportion of variation in across-census tract diabetes prevalence within NYC, even when model fitting was blind to NYC data. Our results add to the literature by quantifying the total predictive

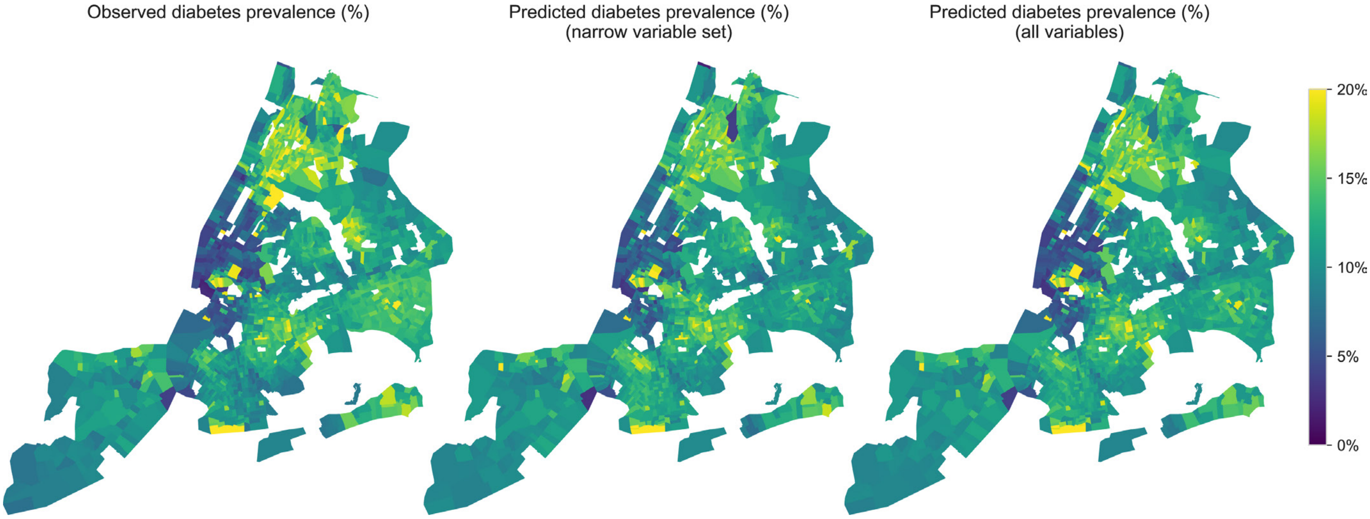
Observed diabetes prevalence (%)
Predicted diabetes prevalence (%)
(narrow variable set)
Predicted diabetes prevalence (%)
(all variables)
20%
15%
10%
5%
0%

Figure 2 Maps of observed and predicted diabetes prevalence within New York City by census tract. Observed prevalence (left), predicted prevalence from narrow model with 16 SDOH predictors (centre) and predicted prevalence from all-variables model with 81 SDOH predictors (right). SDOH, social determinants of health.

Table 2 SHAP values for New York City (test dataset) from narrow 16 variable model

Variable	Raw mean absolute SHAP value	Normalised mean absolute SHAP value	Cumulative normalised mean absolute SHAP value
% high school education or less	1.20	0.21	0.21
% SNAP benefits	1.06	0.19	0.40
% households w/ broadband	0.68	0.12	0.52
% over 65 living alone	0.60	0.10	0.62
% commute <15 min	0.31	0.06	0.68
EPA National walkability index	0.26	0.05	0.72
% single parent families	0.20	0.04	0.76
% housing w/o complete kitchen	0.19	0.03	0.79
% uninsured	0.18	0.03	0.82
% limited English	0.18	0.03	0.85
% essential labour force	0.18	0.03	0.89
% unemployment	0.18	0.03	0.92
% rent >30% household income	0.14	0.02	0.94
Gini index	0.14	0.02	0.97
Low food access @ half mile	0.14	0.02	0.99
Mean days >EPA PM2.5 standard	0.06	0.01	1.00

EPA, Environmental Protection Agency; PM2.5, particulate matter ≤2.5 µm; SHAP, Shapley Additive Explanations; SNAP, Supplemental Nutrition Assistance Programme.

power that a comprehensive set of SDOH variables might have on diabetes prevalence in NYC, where reducing diabetes rates and impact is a legally mandated priority; we did so by using a flexible, interactive ML approach that can better model the complex associations among SDOH than traditional epidemiological models. As such, our ML models were more accurate than traditional linear models, and they suggested a larger cumulative association between SDOH and diabetes prevalence in NYC than did standard linear statistical methods.

Additionally, our approach identified the most important SDOH driving model estimates of diabetes prevalence within NYC. Although all of our SDOH concepts have shown associations with diabetes in prior work, rates of low educational attainment, high participation in SNAP and low household broadband connectivity were most strongly related to model predictions of higher diabetes prevalence in NYC. As these three variables were highly intercorrelated, a general factor of socioeconomic deprivation is likely a primary predictor of disparities in diabetes prevalence, over and above the other individual SDOH measures in our models. Moreover, as we assessed model performance using data unseen during training, where training data was drawn from geographical areas outside of NYC, our performance estimates do not reflect overfitting to the training data or bias from spatial autocorrelation within NYC. As such, their robustness underscores the pervasive and systematic influence of SDOH on chronic illnesses like diabetes.

Our findings support the well-documented association between SDOH and chronic illnesses like diabetes[14 38]; they also provide a quantifiable assessment of their cumulative influence at a neighbourhood level within NYC. Place-based methods and models like ours can be used by policy-makers to identify modifiable SDOH that might have the largest impact on improving local public health. For example, interventions targeting wealth and education inequalities, the two SDOH suggested by our models as showing the strongest association with diabetes prevalence in NYC, have been shown to have long-term positive consequences on health behaviours and health outcomes for some conditions.[10] However, as of yet, there is little concrete evidence identifying the potential impacts of such interventions on diabetes,[14] suggesting that reducing diabetes disparities could be an important target for future studies of policy change. Nonetheless, these findings should be contextualised within a prioritisation framework. While a recent study found that wealth redistribution would be the quickest way to narrow longevity disparities between the USA and other developed nations,[43] this approach is unlikely to be successful in a capitalistic democracy like the USA.[44] However, there is increasing evidence that improving local economic conditions improves health outcomes.[45 46]

Our study has several limitations. First, observed variables may represent latent constructs; an individual variable's correlation with model predictions may, therefore, be driven by a latent factor and not the variable as defined in the data source or the SDOH literature. For example, some have proposed broadband access as a unique SDOH pillar, promoting health through access to information and telehealth, as well

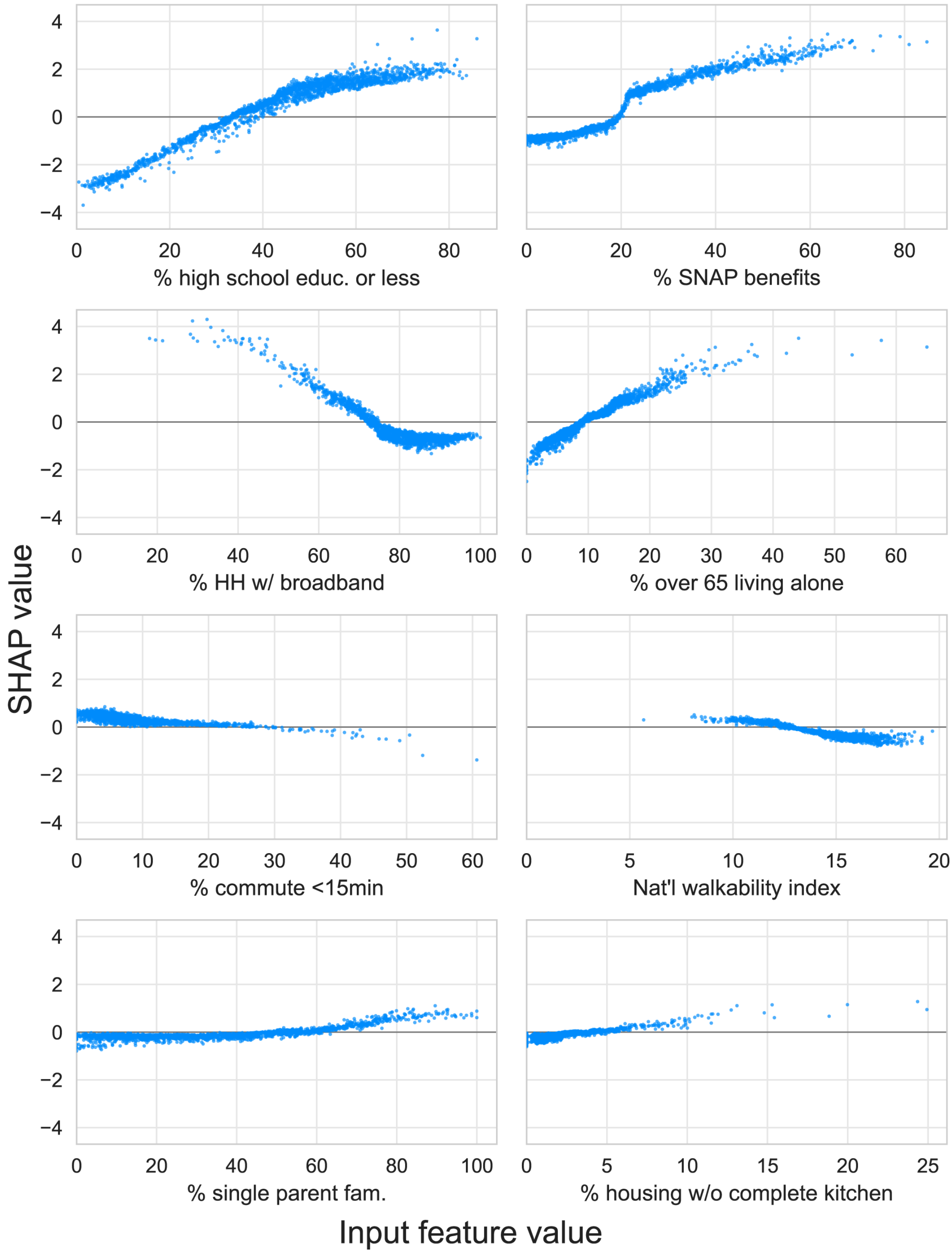

Figure 3 Scatterplots for relationship between SDOH predictors and Shapley Additive Explanations (SHAP) values for New York City (NYC). Top-eight most influential variables are shown. Each point represents one NYC census tract. SHAP values above 0 reflect predictions of diabetes prevalence higher than the mean in percentage points; SHAP values below 0 reflect predictions of diabetes prevalence below the mean in percentage points. SDOH, social determinants of health.

as by modulating other SDOH.[47–50] Although we found low household broadband rates to correlate with higher predicted diabetes prevalence, lack of broadband is generally associated with poverty[51] and was highly correlated with low educational attainment and participation in SNAP in our own data. It is therefore likely that lack of household broadband may simply be a flag for general socioeconomic deprivation, particularly given that broadband is nearly ubiquitously available in all areas of NYC.[52] Second, SHAP values and the relationships seen in the scatterplots (figure 3) cannot be interpreted identically to partial effects in regression models. XGBoost models include higher-order interactions, so that SHAP values for a given variable consider all interactions with other variables in the fitted models. Also, the XGBoost algorithm randomly samples variables between trees, leading to shared predictive power among correlated variables. As such, mean absolute SHAP values and their relative rankings should be interpreted as approximate guides for which variables contribute most to model predictions and not as exact causal contributions.

Third, the relative feature importance hierarchy seen in our test SHAP values applies to NYC and may not be the same in other areas. For example, our measure of poor air quality had low observed variance in NYC compared with the training data (online supplemental figure 3); low variance in the test set could lead to low relative feature importance for NYC. Training or inferencing models in different geographies or with other variable sets could lead to different relative importance. This highlights the importance of our place-based approach in providing local insights. Fourth, our work shares limitations common to all observational research using publicly available survey data. For example, diabetes prevalence was self-reported and not confirmed with medical records, and it does not consider controlled vs uncontrolled diabetes. Additionally, there is the possibility for measurement error, data quality is limited by methods and sampling adequacy concerns tied to BRFSS and other similar data, and survey questions are predefined and outcomes must be interpreted in light of how questions are posed. Fifth, as the level of measurement in our dataset is the census tract, our results might not generalise to individuals. Sixth, our study uses prevalence data rather than incidence data; this has implications for interpreting the relationship between SDOH measures and diabetes. For example, our cross-sectional prevalence data obscures the temporal sequence of events, making it challenging to identify whether specific SDOH drive new cases (incidence) or are associated with existing cases; the relationship between some SDOH variables and new case rates may therefore be different from that reported here. Finally, as our research was observational, our findings, including SHAP importances, are associative and not causal.

Despite these limitations, our interpretable ML findings highlight the substantial impact that SDOH, independent of their associated demographics and health behaviours, can have on chronic disease, such as diabetes. Timely, accurate and high-quality data are a critical component of public health decision-making, and we have shown that ML can support this. These place-based insights can be of practical use, for instance, helping to guide the targeted development of a diabetes incidence reduction plan, as required by NYC law,[21] and similar methods can be adopted by other jurisdictions to guide their own targeted efforts. However, any interventions that arise from this or similar work should target latent factors giving rise to the most impactful SDOH variables identified by our models. As low socioeconomic status and low educational attainment were the most influential variables in our models, we hypothesise that interventions targeting these underlying general determinants may have more long-term impact than interventions focused narrowly and exclusively on one particular indicator of them.

Contributors Study concept and design: DT, YZ, JEC, PS, EA, AA, JMLF and WBW. Acquisition, analysis or interpretation of data: DT, YZ, JEC, PS, EA, AA, JMLF and WBW. Drafting of the manuscript: DT, YZ and WBW. Criticial review of the manuscript for important intellectual content: DT, YZ, JEC, PS, EA, AA, JMLF and WBW. Statistical analysis: DT. Administrative, technical or material support: JMLF. Supervision: WBW. Guarantor: DT.

Funding YZ and JEC received funding for this work from the Novartis Foundation (YZ: NF 231850-01; JEC: NF 100101154).

Disclaimer PS, EA and AA, as employees of the Novartis Foundation, were involved in the design and conduct of the study, interpretation of the results, review and approval of the final manuscript and decision to submit the manuscript for publication. No employees from the Novartis Foundation were involved in the drafting of the manuscript or statistical analysis. The entire research team, including those employed by the Novartis Foundation, had full autonomy in all aspects of this study. The Novartis Foundation did not have power to prevent submission of the manuscript for publication.

Map disclaimer The depiction of boundaries on this map does not imply the expression of any opinion whatsoever on the part of BMJ (or any member of its group) concerning the legal status of any country, territory, jurisdiction or area or of its authorities. This map is provided without any warranty of any kind, either express or implied.

Competing interests None declared.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Ethics approval This study involves human participants. As all data were publicly available, IRB approval was not required. This study uses publicly available data that are aggregated at the census tract level. Diabetes prevalence data were obtained from the US CDC PLACES database, which are publicly available for anyone to download. SDOH variables were obtained from a range of US government agencies, again all publicly available for anyone to download.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data are freely and publicly available for download. Diabetes prevalence data are available from the CDC PLACES data portal at https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/4ai3-zynv/about_data. AHRQ data are available at https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html. USDA Food Access Research Atlas data are available at https://www.ers.usda.gov/data-products/food-access-research-atlas/. CDC Agency for Toxic Substances Disease Registry 2022 Environmental Justice Index are available at https://www.atsdr.cdc.gov/placeandhealth/eji/index.html.

ORCID iD

Darren Tanner http://orcid.org/0000-0003-0805-0494

REFERENCES

1 World Health Organization. Social determinants of health. Available: https://www.who.int/health-topics/social-determinants-of-health [Accessed 07 Mar 2024].

2 Artiga S, Hinton E. Beyond health care: the role of social determinants in promoting health and health equity. Kais. Fam. Found; 2018. Available: https://www.kff.org/racial-equity-and-health-policy/issue-brief/beyond-health-care-the-role-of-social-determinants-in-promoting-health-and-health-equity/ [Accessed 07 Mar 2024].

3 Singh, PhD GK, Daus, MA GP, Allender, MS, BSN, RN M, *et al*. Social Determinants of Health in the United States: Addressing Major Health Inequality Trends for the Nation, 1935-2016. *Int J MCH AIDS* 2017;6:139–64.

4 Costa DL. Health and the Economy in the United States, from 1750 to the Present. *J Econ Lit* 2015;53:503–70.

5 Whitman A, DeN, Chappel A, *et al*. *Addressing Social Determinants of Health: Examples of Successful Evidence-Based Strategies and Current Federal Efforts*. Washington, DC: US Department of Health and Human Services, Office of Health Policy, 2022.

6 National Academies of Sciences Engineering, Medicine. *Communities in Action: Pathways to Health Equity*. Washington, DC: The National Academies Press, 2017.

7 US Centers for Disease Control and Prevention. Health in all policies. 2023. Available: https://www.cdc.gov/policy/hiap/index.html [Accessed 07 Mar 2024].

8 White House. The biden-harris administration immediate priorities. Available: https://www.whitehouse.gov/priorities/ [Accessed 07 Mar 2024].

9 Butkus R, Rapp K, Cooney TG, *et al*. Envisioning a Better U.S. Health Care System for All: Reducing Barriers to Care and Addressing Social Determinants of Health. *Ann Intern Med* 2020;172:S50–9.

10 Thornton RLJ, Glover CM, Cené CW, *et al*. Evaluating Strategies For Reducing Health Disparities By Addressing The Social Determinants Of Health. *Health Aff (Millwood)* 2016;35:1416–23.

11 US Centers for Disease Control and Prevention. Heart disease prevalence. 2023. Available: https://www.cdc.gov/nchs/hus/topics/heart-disease-prevalence.htm [Accessed 07 Mar 2024].

12 World Health Organization. Cardiovascular diseases (CVDs). Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) [Accessed 07 Mar 2024].

13 American Heart Association. Cardiovascular disease and diabetes. Available: https://www.heart.org/en/health-topics/diabetes/diabetes-complications-and-risks/cardiovascular-disease--diabetes [Accessed 07 Mar 2024].

14 Hill-Briggs F, Adler NE, Berkowitz SA, *et al*. Social Determinants of Health and Diabetes: A Scientific Review. *Diabetes Care* 2021;44:258–79.

15 Hill-Briggs F. 2018 Health Care & Education Presidential Address: The American Diabetes Association in the Era of Health Care Transformation. *Diabetes Care* 2019;42:352–8.

16 Golden SH, Maruthur N, Mathioudakis N, *et al*. The Case for Diabetes Population Health Improvement: Evidence-Based Programming for Population Outcomes in Diabetes. *Curr Diab Rep* 2017;17:51.

17 Haire-Joshu D, Hill-Briggs F. The Next Generation of Diabetes Translation: A Path to Health Equity. *Annu Rev Public Health* 2019;40:391–410.

18 McNeill E, Lindenfeld Z, Mostafa L, *et al*. Uses of Social Determinants of Health Data to Address Cardiovascular Disease and Health Equity: A Scoping Review. *J Am Heart Assoc* 2023;12:e030571.

19 Vo A, Tao Y, Li Y, *et al*. The Association Between Social Determinants of Health and Population Health Outcomes: Ecological Analysis. *JMIR Public Health Surveil* 2023;9:e44070.

20 Platkin C, Nielsen A. Testimony about NYC's efforts to address the diabetes epidemic. NYC Food Policy Cent. Hunt. Coll; 2023. Available: https://www.nycfoodpolicy.org/testimony-to-the-new-york-city-council-on-efforts-to-address-the-growing-diabetes-epidemic/ [Accessed 22 Mar 2024].

21 Requiring the department of health and mental hygiene to develop and implement a citywide diabetes incidence and impact reduction plan. Available: https://nyc.legistar.com/LegislationDetail.aspx?ID=6012785&GUID=FEBE7CA3-DFCA-4FF5-BFEB-7DCD946E46AF [Accessed 29 Mar 2024].

22 Health department launches new center to enhance the agency's data and analytics capacity. NYC Health; 2023. Available: https://www.nyc.gov/site/doh/about/press/pr2023/health-department-launches-new-center-to-enhance-agency-data-analytics.page [Accessed 02 Apr 2024].

23 New york city council passes legislation to improve health and extend life expectancy for all New Yorkers. N. Y. City Counc; 2024. Available: https://council.nyc.gov/press/2024/02/08/2558/ [Accessed 02 Apr 2024].

24 Novartis Foundation. AI4HealthyCities. Available: https://www.novartisfoundation.org/transforming-population-health/ai4healthycities [Accessed 07 Mar 2024].

25 Stevens LM, Mortazavi BJ, Deo RC, *et al*. Recommendations for Reporting Machine Learning Analyses in Clinical Research. *Circ Cardiovasc Qual Outcomes* 2020;13:e006556.

26 Molnar C. *Interpretable Machine Learning*. 2023.

27 Greenlund KJ, Lu H, Wang Y, *et al*. PLACES: Local Data for Better Health. *Prev Chronic Dis* 2022;19:E31.

28 Centers for Disease Control and Prevention. PLACES: local data for better health, census tract data 2020 release. Available: https://data.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-Census-Tract-D/4ai3-zynv/about_data [Accessed 22 Jun 2023].

29 Centers for disease control and prevention behavioral risk factor surveillance system. Available: https://www.cdc.gov/brfss/about/index.htm [Accessed 22 Feb 2024].

30 CDC. 500 cities project: 2016 to 2019 | places: local data for better health. Available: https://www.cdc.gov/places/about/500-cities-2016-2019/index.html [Accessed 22 Feb 2024].

31 Centers for disease control and prevention 500 cities: local data for better health, 2018 release. Available: https://data.cdc.gov/500-Cities-Places/500-Cities-Local-Data-for-Better-Health-2018-relea/rja3-32tc/about_data [Accessed 22 Jun 2023].

32 Agency for Healthcare Research and Quality. Social determinants of health database. Available: https://www.ahrq.gov/sdoh/data-analytics/sdoh-data.html [Accessed 29 May 2023].

33 CDC, Centers for Disease Control and Prevention and Agency for Toxic Substances Disease Registry. Environmental justice index. 2022. Available: https://www.atsdr.cdc.gov/placeandhealth/eji/index.html [Accessed 29 May 2023].

34 US department of agriculture food access research atlas. 2019. Available: https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/ [Accessed 29 May 2023].

35 Flanagin A, Pirracchio R, Khera R, *et al*. Reporting Use of AI in Research and Scholarly Publication-JAMA Network Guidance. *JAMA* 2024;331:1096–8.

36 Brown AF, Ettner SL, Piette J, *et al*. Socioeconomic position and health among persons with diabetes mellitus: a conceptual framework and review of the literature. *Epidemiol Rev* 2004;26:63–77.

37 Walker RJ, Strom Williams J, Egede LE. Influence of Race, Ethnicity and Social Determinants of Health on Diabetes Outcomes. *Am J Med Sci* 2016;351:366–73.

38 Hill-Briggs F, Fitzpatrick SL. Overview of Social Determinants of Health in the Development of Diabetes. *Diabetes Care* 2023;46:1590–8.

39 Walker RJ, Smalls BL, Campbell JA, *et al*. Impact of social determinants of health on outcomes for type 2 diabetes: a systematic review. *Endocrine* 2014;47:29–48.

40 Centers for Disease Control and Prevention. PLACES Health Outcomes Measure Definitions. PLACES Health Outcomes Meas. Defin. 2023. Available: https://www.cdc.gov/places/measure-definitions/health-outcomes/index.html [Accessed 25 Mar 2024].

41 Chen T, Guestrin C. XGBoost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016:785–94.

42 Lundberg SM, Lee S-I, *et al*. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, eds. *Advances in Neural Information Processing Systems*. Curran Associates, 2017.

43 Himmelstein KEW, Tsai AC, Venkataramani AS. Wealth Redistribution to Extend Longevity in the US. *JAMA Intern Med* 2024;184:311–20.

44 Weeks WB, Lavista Ferres JM, Weinstein JN. Health and Wealth in America. *Int J Public Health* 2024;69:1607224.

45 Khatana SAM, Venkataramani AS, Nathan AS, *et al*. Association Between County-Level Change in Economic Prosperity and Change in Cardiovascular Mortality Among Middle-aged US Adults. *JAMA* 2021;325:445–53.

46 Wallace HOW, Fikri K, Weinstein JN, *et al*. Improving Economic Conditions Matter for Mortality: Changes in Local Economic Distress Associated with Mortality Among Medicare Fee-for-Service Beneficiaries Between 2003 and 2015. *J Gen Intern Med* 2022;37:249–51.

47 Benda NC, Veinot TC, Sieck CJ, *et al*. Broadband Internet Access Is a Social Determinant of Health! *Am J Public Health* 2020;110:1123–5.

48 Bauerly BC, McCord RF, Hulkower R, *et al*. Broadband Access as a Public Health Issue: The Role of Law in Expanding Broadband Access and Connecting Underserved Communities for Better Health Outcomes. *J Law Med Ethics* 2019;47:39–42.

49 Early J, Hernandez A. Digital Disenfranchisement and COVID-19: Broadband Internet Access as a Social Determinant of Health. *Health Promot Pract* 2021;22:605–10.

50 Kickbusch I, Piselli D, Agrawal A, *et al*. The Lancet and Financial Times Commission on governing health futures 2030: growing up in a digital world. *The Lancet* 2021;398:1727–76.

51 Swenson K, Ghertner R. *People in Low-Income Households Have Less Access to Internet Services*. U.S. Department of Health & Human Services, 2020.

52 New york state psc broadband map. Available: https://mapmybroadband.dps.ny.gov/ [Accessed 05 Apr 2024].