How Food Environment Impacts Dietary Consumption and Body Weight: A Country-wide Observational Study of 1.5 Billion Food Logs

Tim Althoff [1], PhD; Hamed Nilforoshan[2], B; Jenna Hua[3], RD, MPH, PhD; Jure Leskovec[4], PhD

[1] Department of Computer Science, Stanford University
[2] Department of Computer Science, Columbia University
[3] Department of Health Research & Policy, Stanford University School of Medicine
[4] Department of Computer Science, Stanford University

Corresponding Author: Tim Althoff, PhD, Department of Computer Science, Stanford University, William Gates Building, Room 414, Stanford, CA 94305-9040 (althoff@cs.stanford.edu)

Key Points

Question
What is the impact of food access, income and education on food consumption and weight status?

Findings
In this observational study that included 1,050,493 subjects across 9815 U.S. zip codes, higher access to grocery stores, lower access to fast food, higher income and education were independently associated with higher consumption of fresh fruits and vegetables, lower consumption of fast food and soda, and lower body mass index, but these associations varied significantly across black, hispanic and white populations.

Meaning
Policy targeted at improving food access, income and education may increase healthy eating, but interventions may need to be targeted to specific subpopulations for optimal effectiveness.

Abstract

IMPORTANCE

An unhealthy diet is a key risk factor for chronic diseases including obesity, diabetes, and heart disease. Limited access to healthy food options may contribute to unhealthy diets. However, previous studies of food environment have led to mixed results, potentially due to methodological limitations of small sample size, single location, and non-uniform methodology across studies.

OBJECTIVE

To quantify the independent impact of fast food and grocery access, income and education on food consumption and weight status.

DESIGN, SETTING AND PARTICIPANTS

Retrospective cohort study of 1,050,493 subjects across 9815 U.S. zip codes logging 1.5 billion consumed foods. Participants were users of the MyFitnessPal smartphone application and used the app to monitor their caloric intake for an average of 197 days each (min 1, max 2756 days).

MAIN OUTCOMES AND MEASURES

The primary outcomes were relative changes in consumption of fresh fruits and vegetables, fast food, and soda, as well as relative change in body mass index (BMI). Food consumption logs were classified into categories including fresh fruits and vegetables, fast food and soda, and these measures were aggregated on the zip code level. Food access measures for each zip code were computed from USDA Food Access Research Atlas and Yelp sources, and demographic, income and education measures were based on Census data. To assess the independent impact of food access, income and education, propensity score matching methods were used to create pairs of zip codes that vary in one variable and are very similar in all other variables and in terms of race and ethnicity.

RESULTS

- Among 1,050,493 participants across 9815 U.S. zip codes, access to grocery stores, non-fast food restaurants, income and education were independently associated with healthier food consumption and lower body mass index.
- Specifically, in zip codes of with above median household income, users logged 4.6% more F&V, 8.8% less fast food, 5.3% less soda and were 4.5?% less likely to be overweight or obese, relative to matched below median household income zip codes (all $P<0.05$; Bootstrap resampling).
- In zip codes of above median education users logged 4.7% more F&V, 6.1% less fast food, 0.8% less soda and were 8.5?% less likely to be overweight or obese (all $P<0.05$ except soda

consumption; Bootstrap resampling).

- In zip codes of high grocery access users logged 5.6% more F&V, 10.8% less fast food, 8.0% less soda and were 3.8?% less likely to be overweight or obese (all $P<0.05$; Bootstrap resampling).
- In zip codes of low fast food access users logged 4.3% more F&V, 8.0% less fast food, 8.0% less soda and were 2.3?% less likely to be overweight or obese (all $P<0.05$; Bootstrap resampling).
- However, substantial differences were observed between predominantly black, hispanic, and white zip codes. For instance, within predominantly black black zip codes we found no significant impact of higher income on food consumption or weight status. Further, within predominantly hispanic zip codes high grocery access, in contrast to high income, high education, and low fast food access, was the only factor associated with higher consumption of fresh fruits and and vegetables.

CONCLUSIONS AND RELEVANCE

Policy targeted at improving access to grocery stores, access to non-fast food restaurants, income and education may increase healthy eating, but interventions may need to be adapted based on specific subpopulation (black, hispanic, or white) for optimal effectiveness.

Introduction

According to the Global Burden of Disease Study and related studies, unhealthy diet generates a bigger non-communicable disease (NCD) burden than tobacco, alcohol and physical inactivity combined, and is the leading risk for death and disability globally [1]. Improving diet as a means to target NCDs, such as obesity, hypertension, heart disease, stroke, diabetes, kidney disease, or cancer, is essential [2].

The obesity epidemic is a key contributor to the global burden of chronic disease and disability, with obesity and overweight being major risk factors for Type 2 diabetes, cardiovascular disease, and several cancers [3]. In the United States, an estimated 70.7% of adults age 20 years have overweight or obesity. Because obesity is a major driver of diabetes, cardiovascular diseases, and cancers, it has been estimated that even a one percentage point reduction from the predicted trend to 2030 would reduce obesity-caussed medical expenditures by $84.9 billion over two decades [4]. The focus of obesity etiologic research has been on energy imbalance, particularly diet and physical activity behaviors. However, emerging evidence suggest that upstream factors such as the built environment/ food environment, behavioral, and socioeconomic cues and triggers have been proposed to influence energy intake and/or expenditure, significantly affect diet and physical activity [6].

Despite many obesity risk studies focused on energy balance, and the associations between obesity and food environments, most have been constrained by deficiencies in current methodologies for assessing personal dietary patterns and food environment exposures. Additionally, many food environment studies have overlooked the importance of personal perceptions on availability, accessibility, affordability, accommodation and acceptability of the food environment [7]. Furthermore, recent studies have produced mixed results on environmental effects on eating, potentially due to small and single-location populations and methodological differences across studies [8].

Most of the above challenges can be addressed with the advances in smartphone and personal sensing technologies. With the advent of ubiquitous behavioral sensing through smartphones and wearable devices and the development of "big data" analytics,personal diet patterns can be tracked continuously in real-time using the smartphone. Additionally, with the availability of large geospatial data that can be mined through Google and other Internet sources, there are now unprecedented opportunities to combine the locational data on diets tracked with smartphones with food establishment data in many places in the world. There is accelerated use of smartphones worldwide, with the adoption rate among adults currently at ~69% in developed countries (almost double what it was 5 years ago), and 46% in developing economies and growing rapidly [9]. Of note, U.S. subgroups who rely on smartphones for online access at elevated levels

include those with low household incomes and educational attainment, as well as racial/ethnic minority groups (in particular, African Americans and Latinos).

Methods

Study Design and Population

We conducted a United States countrywide cross-sectional study of participants' self-reported food intake and body-mass index (BMI) in relation to demographic (education, ethnicity), socioeconomic (income), and food environment factors(grocery store and fast food access) captured on zip-code level. Objective measures of these factors discussed in more detail below.

Overall, this study analyzed 1.5 billion food intake logs from 1,050,493 U.S. smartphone users over seven years across 9815 zip codes (US has total of 41,685 zip codes). Participants were users of the MyFitnessPal app, a free application for tracking caloric intake. We analyzed anonymized, retrospective data collected during a 7-year observation period between 2010 and 2016 that were aggregated to the zip code level. able includes basic statistics on study population demographics and weight status (Body Mass Index; BMI). Data handling and analysis was conducted in accordance with the guidelines of the Stanford University Institutional Review Board.

	Overweight	Obese	Median Age (years)	Gender	Body Mass Index	Median Family Income per Zipcode
Our Study	34.7%	33.7%	36.0	74.2% female	28.60	$67,322
National Average	32.8%	37.9%	37.7	50.5% female	26.5	$59,039

Demographic and socioeconomic measures

We obtained data on demographic and socioeconomic factors from CensusReporter (https://censusreporter.org/). Specifically, for each zip code in our data set we obtained median family income (2010-14 American Community Survey's census tract estimates [1below]), fraction of population with college education (Bachelor's degree or higher), and fraction of population that is black, hispanic, or white.

Food environment measures

We obtained data on grocery store access (fraction of population that is more than 0.5 miles away from grocery store) and food desert status from the USDA Food Access Research Atlas. A census tract was considered a food desert if at least 500 people or 30 percent of residents live more than 1 mile from a supermarket in urban areas (10 miles in rural areas). We aggregated these data on census tract level to the zip code level by requiring that 100% of census tracts included in a given zip codes are food deserts, and by taking average of each census tract grocery store access measure, weighted by the number of people in the tract.

We measured fast food access through the fraction of restaurants that are fast food restaurants within a 40 km (25 miles) radius of the zip code center. Data on local restaurants and businesses were obtained through the Yelp API. For each zip code, we consider up to 1000 restaurant businesses that are nearest to the zip code center up to a distance of 40km (67.8% of zip code queries resulted in 1000 restaurant businesses within 40km; Yelp API results are restricted to 1000 results). Zip codes with more than 1000 restaurant businesses within 40km are likely more dense urban environments and we still capture the nearest restaurant businesses to the zip code center that arguably capture the local food environment. .

Outcome measure (food consumption and weight status)

We used 1.5 billion food intake logs by 1,050,493 U.S. users of the MyFitnessPal smartphone application to quantify food consumption across 9815 zip codes. During the observation period from June 2nd, 2005 to November 15th, 2016, the average user logged 10.06 entries into their digital food journal per day. The average user used the app for 197 days. Days. Almost users in this sample (99.9%) used the app for at least 8 days.

Each entry contains a separate food component (e.g., banana, yoghurt, etc.). We classified all entries into three categories: fresh fruits and vegetables (F&V; through a proprietary classifier by MyFitnessPal), fast food (if the description contained the name of a fast food chain), and soda (if the description contained the name of a soda drink). In all cases, descriptions were normalized by lowercasing and removing punctuation. We then aggregated all food entries for each zip code and computed the fraction of entries that were in the F&V, fast food, and soda categories. We further used the average body-mass index (BMI) in each zip code as a weight status outcome. BMI was self-reported by users of the smartphone application

Data Validation

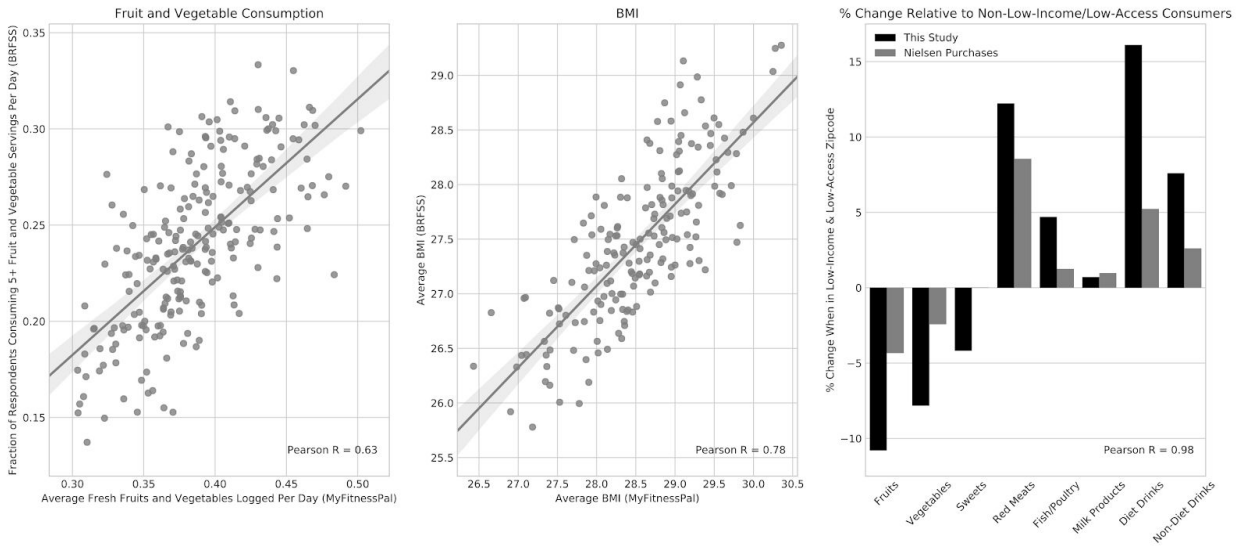

Figure 1: *Smartphone-based food logs correlated with previous, representative survey measures and purchase data.*

(a) Fraction of fresh fruits and vegetables logged is correlated with BRFSS survey data (R=0.63, p<10^-5; Methods). (b) Body Mass Index of smartphone cohort is correlated with BRFSS survey data (R=0.78, p<10^-5; Methods). (c) Digital food logs replicate previous findings of relative consumption differences in low-income, low-access food deserts based on Nielsen purchase data (R=0.93, p<10^-3; Methods).

We find that our study population has significant overlap with the U.S. national population (Demographics table 1) but is skewed towards women and higher income. We demonstrate below that food consumption measured based on this population are highly correlated with state-of-the-art measures.

Smartphone apps such as MyFitnessPal feature large databases with nutritional information and can be used to track one's diet over time. Previous studies have compared app-reported diet measures to traditional measures including 24 h dietary recalls and food composition tables. These studies found that both measures tend to be highly correlated [1,2], but that app-reported measures tend to underestimate certain macro- and micronutrients [1,2], especially in populations that were previously unfamiliar with the smartphone applications [3]. In contrast, this study leverages a convenience sample of existing users of the smartphone app MyFitnessPal.

Yelp data has been used in measures of food environment [4] and a study in Detroit found Yelp data to be more accurate than commercially-available databases [5].

This study uses a combination of MFP data to capture food consumption, Yelp and USDA data to capture food environment, and Census data to capture basic demographics. As a preliminary, basic test, we investigated correlations between the Mexican food consumption, the fraction of Mexican restaurants, and the fraction of Hispanics in the population, on a zip code level. We found

that consumption was correlated with the fraction of Mexican restaurants (R=0.69; P<10^-4) and the fraction of Hispanics in the population (R=0.54; P<10^-4). Further, the fraction of Mexican restaurants was correlated with the fraction of Hispanics in the population as well (R=0.56; P<10^-4).

Reproducing state-of-the-art measures using population-scale digital food logs

To investigate the applicability of population-scale digital food logs to study the impact of food environment, income and education on food consumption, we measured the correlation between our smartphone app-based measures and state-of-the-art measures of food consumption including the Behavioral Risk Factor Surveillance System (BRFSS), based on representative surveys, and the Nielsen Homescan data, which is a commercial food purchase database.

We used the latest available survey data from BRFSS (https://www.cdc.gov/brfss/). Specifically, we used variables FV5SRV (BRFSS 2009) representing the the faction of people eating five or more servings of fresh fruit and vegetables, and BMI5 representing body mass index (BRFSS 2012). Comparing to BRFSS on a county level, the average number of F&V logged per day (MFP) was correlated with the fraction of respondents that report consumptions of at least five servings of F&V per day (R=0.63, P<.0001). Further, average county-level BMI was strongly correlated as well (R=0.78, P<.0001). We further compared to published results by the USDA, which used data from the 2010 Nielsen Homescan Panel Survey that captured household food purchases for in-home consumption (but not restaurants and fast food purchases).

We attempted to reproduce published findings on the differences in low-income, low-access communities (food deserts) compared to non-low-income, non-low-access communities across categories of fruit, vegetable, sweets, red meat, fish/poultry, milk products, diet drinks, and non-diet drinks . We used proprietary MFP classifiers to categorize foods logged into these categories. We found that our app-based food logs were very highly correlated with previously published results (R=0.98, P<.0001) and that the absolute differences between food deserts and non-food deserts were stronger in the MFP data compared to Nielsen purchase data.

In total, these results demonstrate convergent validity. Specifically, our results suggest that population-scale digital food logs can reproduce the basic dynamics of traditional, state-of-the-art measures, and they can do so at massive scale and comparatively low cost.

Statistical Analysis

In this large-scale observational study, we used a matching-based approach [10] to disentangle contributions of income, education, grocery access, and fast food access on food consumption. To estimate the impact of each of these factors, we divide all available zip codes into treatment and control groups based on a median split. We then create matched pairs of zip codes by selecting a zip code in the control group that is closely matched across all factors but the current

treatment. In addition, we control for the demographic makeup of each zip code by matching on fraction of black, hispanic, and non-hispanic white populations in the zipcode. Through this process, we attempt to eliminate variation of plausible influences and to isolate the effect of interest. We repeat this process for each treatment of interest; for example for the results presented in Figure 1, we performed four matches, one for each of income, education, grocery access and fast food access.

Specifically, we use a one-to-to nearest neighbor matching approach that uses the Mahalanobis distance, replacement, and a caliper to ensure good balance between treated and control units [10]. The caliper was chosen to ensure good balance between treated and control units, which was defined as a Standardized Mean Difference (SMD) of less than 0.25 standard deviations [10]. Some definitions of SMD use the standard variation in the overall population before matching [10]. However, we choose the standard deviation in the control group post-matching, which typically is much smaller and therefore gives more conservative estimates of balance between treated and control units [10]. For the subpopulation experiments (Figure 3), we use a 3:1 matching that assigns three control units to each treatment in order to minimize variance due to the smaller sample sizes.

Balancing statistics for each of the matches are available in the Appendix. We tested discriminant validity our our statistical approach by measuring the effect of null-treatments that should not have any impact on food consumption. We chose examples of null-treatments by selecting variables that had little correlation with study independent variables (income, education, grocery access, fast food access) and were plausibly unrelated to food consumption. This selection process lead to use of the fraction of communication, IT, and cleaning services nearby as measured through Yelp. Applying our analysis pipeline to these null-treatments, we found effect estimates that were close to zero. This demonstrated that our statistical analysis does not produce measurement that it is not supposed to measure; that is, discriminant validity.

Ethics approval

Data handling and analysis was conducted in accordance with the guidelines of the Stanford University Institutional Review Board.

Results

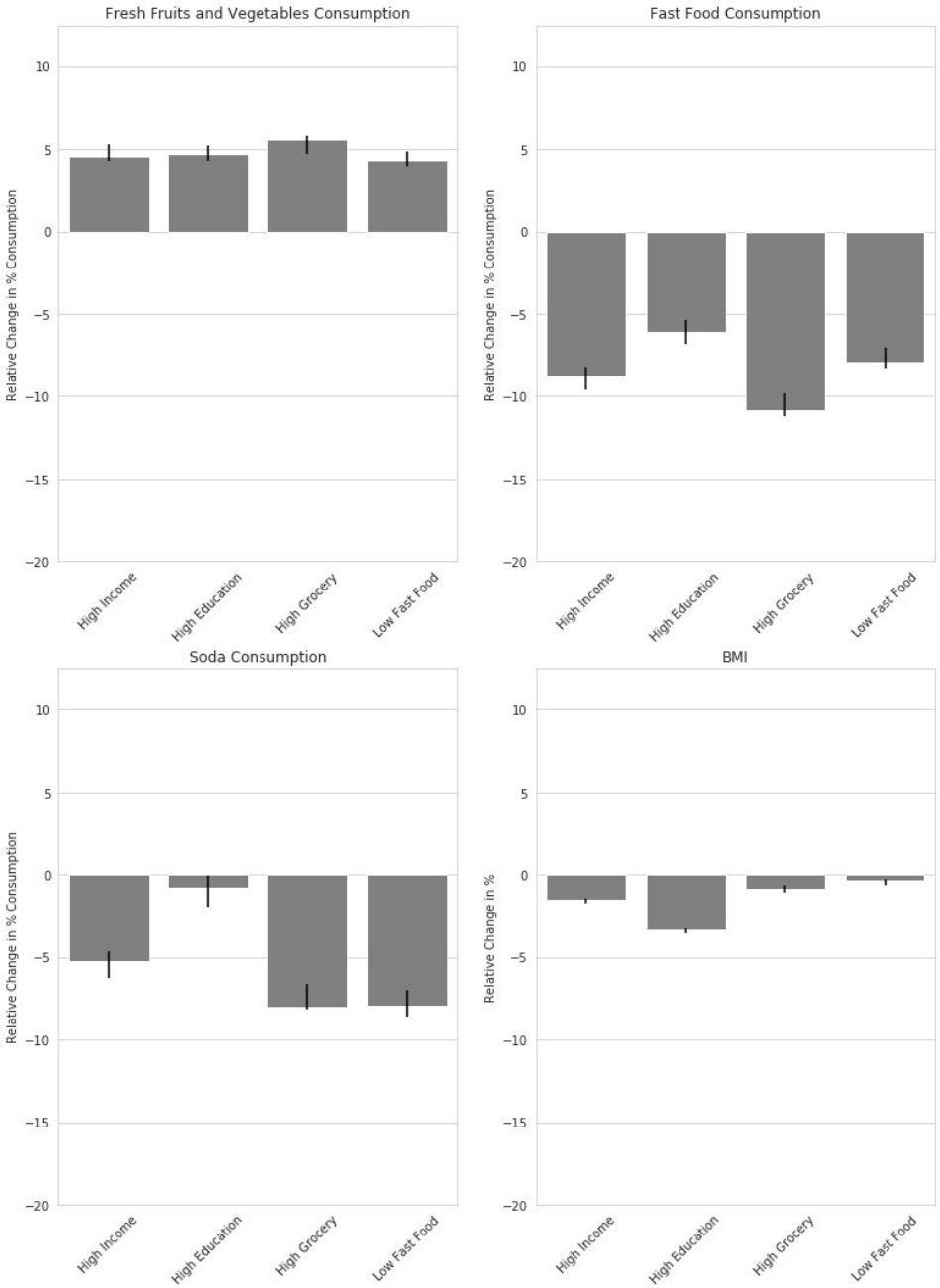

Figure 2: Independent contributions of high income (median family income higher than $67,321), high education (fraction of population with college education 61.3% or higher), high grocery access (fraction of population that is more than 0.5 miles from nearest grocery store is 22.2% or less), and low fast food access (less than 3.6% of all businesses are fast-food chains) on relative change in consumption of (a) fresh fruits and vegetables, (b) fast food, (c) soda, and (d) relative change in body mass index (BMI). Cut points correspond to the median variable value. Estimates are based on matching experiments controlling for all but one treatment variable and distribution of ethnicity and race in zip code (Methods). Error bars correspond to bootstrapped 95% confidence intervals (Methods).

Across all 6580 U.S. codes, we found that high income, high education, high grocery access, and low fast food access were associated with higher consumption of F&V, lower consumption of fast food and soda, and lower prevalence of overweight or obese BMI levels (Figure 2).

- Specifically, in zip codes of high income users logged 4.6% more F&V, 8.8% less fast food, 5.3% less soda and were 4.5?% less likely to be overweight or obese (all $P<0.05$; Bootstrap resampling).
- In zip codes of high education users logged 4.7% more F&V, 6.1% less fast food, 0.8% less soda and were 8.5?% less likely to be overweight or obese (all $P<0.05$ except soda consumption; Bootstrap resampling).
- In zip codes of high grocery access users logged 5.6% more F&V, 10.8% less fast food, 8.0% less soda and were 3.8?% less likely to be overweight or obese (all $P<0.05$; Bootstrap resampling).
- In zip codes of low fast food access users logged 4.3% more F&V, 8.0% less fast food, 8.0% less soda and were 2.3?% less likely to be overweight or obese (all $P<0.05$; Bootstrap resampling).

Note that the reported effect size are based on comparing above and below median zip codes for any given factor. We found increased effect sizes when comparing top versus bottom quartiles (Figure SI1), suggesting a dose-response relationships across all variables. We found that zip codes with high grocery access compared to low grocery access had the largest relative increases in F&V, fast food, and soda (tied) consumption. However, in terms of its impact on overweight and obese BMI levels, we found high education and high income to be more effective. In particular, high education zip codes had the largest relative decrease in overweight and obese BMI levels (8.?%), even though education's impact on soda consumption was insignificant (0.8%).

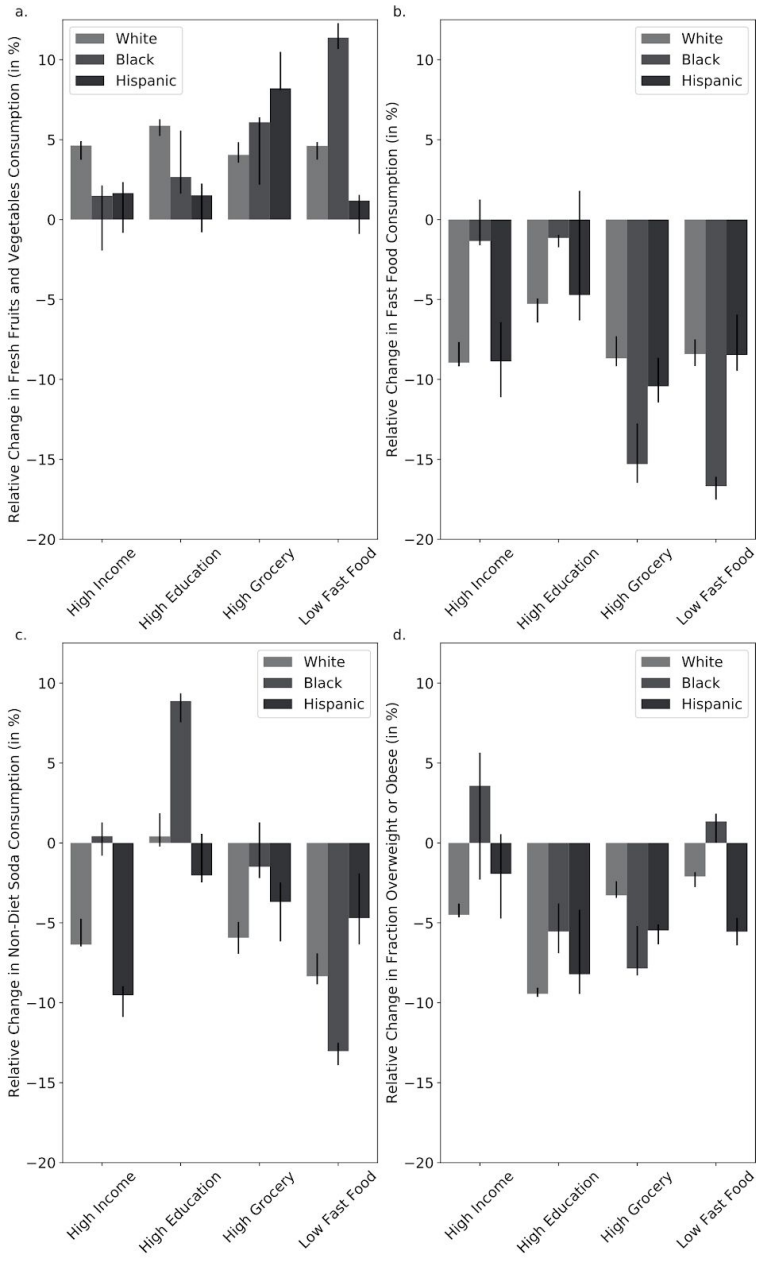

Figure 3: Independent contributions of high income (median family income higher than $67,321), high education (fraction of population with college education 61.3% or higher), high grocery access (fraction of population that is at less than 1/2 mile away from nearest grocery store is 77.8% or higher), and low fast food access (less than 3.6% of all businesses are fast-food chains) on relative change in consumption of (a) fresh fruits and vegetables, (b) fast food, (c) soda, and (d) relative change in body mass index (BMI). **Estimates are stratified by predominantly (i.e. 50% or more) black, hispanic, and non-hispanic white zip codes.** Estimates are based on matching experiments controlling for all but one treatment variable and distribution of ethnicity and race in zip code (Methods). Error bars correspond to bootstrapped 95% confidence intervals (Methods).

We separately repeated our analyses within zip codes that were predominantly black (4.5%), hispanic (6.8%) and non-hispanic white (72%). Results within predominantly non-hispanic white zip codes closely matched results within the overall population. However, restricting our analyses to predominantly black and hispanic zip codes led to a remarkably different findings.

Specifically, within predominantly black zip codes we found no significant impact of higher income across across all outcome variables. The effect of higher education was low compared to other factors in these zip codes, though it led to both 8% lower fraction of overweight and obese BMI levels and 5.3% higher soda consumption. High grocery access had the strongest impact on overweight & obese BMI levels (7% less) and and fast food consumption (16.2% less). Within predominantly black zip codes, low fast food access was associated with the highest increases in F&V consumption (8.2%), the highest decrease in fast food consumption (9.4%), and the highest decrease soda consumption (12.5%), but was not associated with a significant change in BMI levels (1.5?% increase).

In contrast, within predominantly hispanic zip codes we found a significant effect of high income on lower fast food consumption (12.7%) and lower soda consumption (11.6%), but not on F&V consumption (1.0%) and BMI levels (2.0% increase). Higher education zip codes had significantly reduced BMI levels (7% relative decrease), but no other significant associations. High grocery access was the only factor associated with higher F&V consumption (9.4%), and among the most important factors to reduce fast food consumption (9.8%), reduce soda consumption (2.1%), and reduced burden of overweight (5%). Lower fast food access had similar effects, except that the impact on F&V consumption was insignificant (0.6%).

Few factor led to consistent improvements across all three subpopulations. Across all groups, F&V consumption was strongly increased by high grocery access, fast food consumption was most reduced in high grocery access and low fast food access zip codes, soda consumption was most reduced in low fast food environments, and BMI levels were reduced across all groups only in high education zip codes.

Bibliography

[1] Centers for Disease Control and Prevention. "Healthy, Hunger-Free Kids Act of 2010, Section 204: Local School Wellness Policies. 5-Year Technical Assistance and Guidance Plan." (2011).

[2] Andersson, Agneta, and Susanne Bryngelsson. "Towards a healthy diet: from nutrition recommendations to dietary advice." Scandinavian Journal of Food and Nutrition 51.1 (2007): 31-40.

[3] Leong, King Sun, and John P. Wilding. "Obesity and diabetes." Best Practice & Research Clinical Endocrinology & Metabolism 13.2 (1999): 221-237.

[4] Go, Alan S., et al. "Heart disease and stroke statistics—2013 update a report from the American Heart Association." Circulation (2012): CIR-0b013e31828124ad.

[5] Mattes, Richard, and Gary D. Foster. "Food environment and obesity." Obesity 22.12 (2014): 2459-2461.

[6] Cobb, Laura K., et al. "The relationship of the local food environment with obesity: a systematic review of methods, study quality, and results." Obesity 23.7 (2015): 1331-1344.

[7] Whelan, Amanda, et al. "Life in 'food desert'." Urban Studies 39.11 (2002): 2083-2100.

[8] Wrigley, Neil, et al. "Assessing the impact of improved retail access on diet in a'food desert': a preliminary report." Urban Studies 39.11 (2002): 2061-2082.

[9] Park, Yangil, and Jengchung V. Chen. "Acceptance and adoption of the innovative use of smartphone." Industrial Management & Data Systems 107.9 (2007): 1349-1365.

[10] Diamond, Alexis, and Jasjeet S. Sekhon. "Genetic matching for estimating causal effects: A general multivariate matching method for achieving balance in observational studies." Review of Economics and Statistics 95.3 (2013): 932-945.

ARTICLE INFORMATION

Author Affiliations: Department of Computer Science, Stanford University (Althoff, Tim; Leskovec, Jure); Department of Computer Science, Columbia University (Nilforoshan, Hamed); Stanford Prevention Research Center, School of Medicine, Stanford University (Hua, Jenna).

Author Contributions:
Study concept and design: Althoff and Leskovec.
Statistical analysis: Althoff and Nilforoshan.
Interpretation of data: All authors.
Drafting of the manuscript: All authors.
Critical revision of the manuscript for important intellectual content: All authors.

Conflict of Interest Disclosure: None reported.

Roles of Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH or sponsors.