Large-Scale Diet Tracking Data Reveal Disparate Impacts of Food Environment

Tim Althoff,[1*] Hamed Nilforoshan[2], Jenna Hua[3], Jure Leskovec[2,4]

[1]Allen School of Computer Science & Engineering, University of Washington

[2]Department of Computer Science, Stanford University

[3]Stanford Prevention Research Center, Department of Medicine, Stanford University School of Medicine

[4]Chan Zuckerberg Biohub, San Francisco, CA

*To whom correspondence should be addressed;

E-mail: althoff@cs.washington.edu

An unhealthy diet is a key risk factor for chronic diseases including obesity, diabetes, and heart disease[1,2]. Limited access to healthy food options may contribute to unhealthy diets[3,4]. Studying diets is challenging, typically restricted to small sample sizes, single locations, and non-uniform design across studies, and has led to mixed results on the impact of the food environment[5–21]. Here we leverage smartphones to track diet health and weight status in a country-wide observational study of 1,164,926 U.S. participants and 2.3 billion food entries to study the independent contributions of fast food and grocery access, income and education on these outcomes. This study constitutes the largest nationwide study examining the impact of the food environment on diet to date, with 300 times more participants and 4 times more person years of tracking than the Framingham Heart Study. We find that higher access to grocery stores, lower access to fast food, higher income and education are independently associated with higher consumption of fresh fruits and vegetables, lower consumption of fast food and soda, and lower likelihood of being overweight/obese, but that these associations vary significantly across predominantly Black, Hispanic, and White zip codes. For instance, within predominantly Black zip codes we find that high income is associated with a decrease in healthful food consumption patterns across fresh fruits and vegetables (6.5%) and fast food (5.5%). Further, high grocery access has a significantly larger association with increased fruit and vegetable consumption in predominantly Hispanic (7.4% increase) and Black (10.2% increase) zip codes in contrast to predominantly White zip codes (1.7% increase). 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.

1 Introduction

Unhealthy diet is the leading risk factor for disability and mortality globally[1]. Emerging evidence suggests that the built and food environment, behavioral, and socioeconomic cues and triggers significantly affect diet[5]. Prior studies of the food environment and diet have led to mixed results[5–21], and very few used nationally representative samples. These mixed results are potentially due to methodological limitations of small sample size, differences in geographic contexts, study population, and non-uniform measurements of both the food environment and diet across studies. Therefore, research with larger sample size and using improved and consistent methods and measurements is needed[7,22,23]. With ever increasing smartphone ownership in the U.S.[24] and the availability of immense geospatial data, there are now unprecedented opportunities to combine various data on individual diets, population characteristics (gender, race and ethnicity), socioeconomic status (income and education), as well as food environment at large scale. Interrogation of these rich data resources to examine geographical and other forms of heterogeneity in the effect of food environments on health could lead to the development and implementation of cost-effective interventions[25]. Here, we leverage large-scale smartphone-based food journals and combine several Internet data sources to quantify the independent impact of food (grocery and fast food) access, income and education on food consumption and weight status of 1,164,926 subjects across 9,822 U.S. zip codes. This study constitutes the largest nationwide study examining the impact of the food environment on diet to date, with 300 times more participants and 4 times more person years of tracking than the Framingham Heart Study[26].

2 Methods

2.1 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.

Overall, this cross-sectional matching-based study analyzed 2.3 billion food intake logs from U.S. smartphone participants over seven years across 9,822 zip codes (U.S. has total of 41,685 zip codes). Participants were participants 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. Comparing our study population to nationally representative survey data, we found that our study population had significant overlap with the U.S. national population in terms of population demographics, education and weight status (Body Mass Index; BMI), but that it was skewed towards women and higher income (Supplementary Table e1). Our matching-based statistical methodology controls for observed biases between comparison groups in terms of income, education, grocery access, and fast food access (Section 2.4). Data handling and analysis was conducted in accordance with the guidelines of the Stanford University Institutional Review Board.

2.2 Study Data We compute outcome measures of food consumption and weight status from 2.3 billion food intake logs by 1,164,926 U.S. participants of the MyFitnessPal (MFP) smartphone application to quantify food consumption across 9,822 zip codes. During the observation period from January 1, 2010 to November 15, 2016, the average participant logged 9.30 entries into their digital food journal per day. The average participant used the app for 197 days. All participants in this sample used the app for at least 10 days. We classified the 2.3 billion food intake entries into three categories of public health interest, fresh fruits and vegetables (F&V), fast food, and sugary non-diet soda, and excluded them from analysis if they did not match these categories. For more details see Supplementary Information. We intentionally use a cross-sectional rather than longitudinal study design, since fine-grained and large-scale temporal data on changes in the food environment were not available.

We obtained data on demographic and socioeconomic factors from CensusReporter[27]. Specifically, for each zip code in our data set we obtained median family income, fraction of population with college education (Bachelor's degree or higher), and fraction of population that is White (not including Hispanic), Black, or Hispanic from the 2010-14 American Community Survey's census tract estimates[27]. While data were available only on zip code level, previous studies have shown that area-level income measures are meaningful for health outcomes and describe unique socioeconomic inequities.[28]

Grocery store access was defined as the fraction of population that is more than 0.5 miles away from a grocery store following the food desert status definitions from the USDA Food Access Research Atlas[29]. Contrary to the USDA definition, we found evidence that even in *rural* zipcodes, the fraction of population greater than 0.5 miles away from grocery stores has the strongest association with food consumption (compared to 10 and 20 miles away), and thus we used 0.5 miles as the threshold in both rural and urban zipcodes (Supplemental Methods: Details on food environment measures). We measured fast food access through the fraction of restaurants that are fast food restaurants within a sample from Yelp, querying the nearest 1000 businesses from the zip code's center, up to a maximum radius of 40 km (25 miles). See Supplementary Information for details and validation of these objective food environment measures.

We will release all data aggregated at a zipcode level in order to enable validation, follow-up research, and use by policy makers.

2.3 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 of over 350,000 adults in the United States [30,31], and the Nielsen Homescan data [32], which is a nationally representative panel survey of the grocery purchases of 169,000 unique households across the United States, based on UPC records of all consumer packaged goods participants purchased from any outlet. (Figure 3). We compare against BRFSS rather than National Health and Nutrition Examination Survey (NHANES), since BRFSS is significantly larger than NHANES, it is remotely administered matching our study, and it has much better geographical coverage than NHANES and geographical comparisons are central to our study.

Comparing our data to BRFSS on county level, we found strong correlations between the amount of fresh fruits and vegetables (F&V) consumed (Figure 3a, R=0.63, $p < 10^{-5}$) and body mass index (Figure 3b, R=0.78, $p < 10^{-5}$). Comparing to USDA purchase data from the Nielsen Homescan Panel Survey we found that our app-based food logs were very highly correlated with previously published results (Figure 3c, R=0.88, $p < 0.01$) and that the absolute differences between food deserts and non-food deserts were stronger in the MFP data compared to Nielsen purchase data. See Supplementary Information for more details. These results demonstrate convergent validity and 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.

2.4 Statistical Analysis In this large-scale observational study, we used a matching-based approach[33,34] 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; that is, we estimate the difference in outcomes between matched above-median and below-median zip codes. We create matched pairs of zip codes by selecting a zip code in the control group that is closely matched to the zip code in the treatment group across all factors, except the treatment factor of interest. Since we repeat this matching process for each zip code in the treatment group, this approach estimates the Average Treatment Effect on the Treated (ATT). 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 4, we performed four matchings, one for each of income, education, grocery access and fast food access. For the sub-population experiments (Figure 5), we repeated the same method on the subset of the zip codes in which the majority of inhabitants were of a particular racial group. See Supplementary Information for details on the matching approach and detailed statistics that demonstrate that treatment and control groups were well-balanced on observed covariates after matching.

We tested discriminant validity of 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 countertop installers, electronics stores, and waterproofing services nearby as measured through Yelp. Applying our analysis pipeline to these null-treatments, we found that all of these null-treatments had zero effect on food consumption. This demonstrated that our statistical analysis approach did not produce measurements that it was not supposed to measure; that is, discriminant validity (Supplementary Figure e2).

3 Results

Across all 9,822 U.S. codes, we found that high income, high education, high grocery access, and low fast food access were independently associated with higher consumption of fresh fruits and vegetables (F&V), lower consumption of fast food and soda, and lower prevalence of overweight or obese BMI levels (Figure 4).[1] Specifically, in zip codes of above median grocery access participants logged 3.4% more F&V, 7.6% less fast food, 8.6% less soda and were 2.4% less likely to be overweight or obese (all $P < 0.001$). In zip codes of below median fast food access participants logged 5.3% more F&V, 6.2% less fast food, 15.7% less soda and were 1.5% less likely to be overweight or obese (all $P < 0.001$). In zip codes of above median education participants logged 9.2% more F&V, 8.5% less fast food, 10.6% less soda and were 13.1% less likely to be overweight or obese (all $P < 0.001$). Finally, in zip codes of above median household income (referred to as *higher income* below), participants logged 3.3% more F&V, 6.8% less fast food, 5.2% less soda (all $P < 0.001$), but had a 0.6% higher likelihood of being overweight or obese ($P = 0.006$). However, above median household income was associated with a 0.34% decrease in BMI ($P < 0.001$). Note that the reported effect size are based on comparing above and below median zip codes for any given factor. We found a general pattern of consistent but increased effect sizes when comparing top versus bottom quartiles (Supplementary Figure e1), suggesting a dose-response relationships across all considered variables (with the exception of the association between low fast food access and likelihood of overweight or obese BMI levels). We found that zip codes with high education levels compared to low education levels had the largest relative increases in F&V, fast food, and overweight or obese BMI levels. However, in terms of its impact on soda consumption, we found low fast food access to be associated with the largest relative decrease in soda consumption. In particular, high education zip codes had the largest relative decrease in reported soda consumption (15.7%).

We separately repeated our data analyses within zip codes that were predominantly Black (3.7%), Hispanic (5.6%) and non-Hispanic White (78.4%) (Figure 5). Results within predominantly non-Hispanic White zip codes closely matched results within the overall population, since most zip codes in this study were predominantly White (78.4%), not unlike to the overall U.S. population (61.3%)[35]. However, restricting our analyses to predominantly Black and Hispanic zip

[1]The only exception to this pattern was a slight (0.34%) increase in overweight or obese BMI levels associated with income.

codes led to remarkably different findings. Specifically, within predominantly Black zip codes we found an impact of higher income in the inverse direction of the population average and towards low healthful food consumption, across across four out of four outcome variables, resulting in decreased F&V consumption (6.5%), increased fast food consumption (5.5%), and increased likelihood of overweight or obese BMI levels (8.1%). Higher income was also associated with increased soda consumption (11.8%) but was not statistically significant ($P = 0.081$). On the other hand, low fast food access and high education access were generally associated with increased diet health, with low fast food access correlating with the highest decrease in soda consumption (14.4%) and high education with the highest increase in fresh fruit and vegetable consumption (11.2%), although decreased fast food access was harmful to one of the outcome variables. Specifically, decreased fast food access was associated with a slight increase in fraction overweight or obese (3.1%). The only variable that had a positive effect on the health of all outcome variables was increased grocery access, which was associated with increased F&V consumption (10.2%), decreased fast food consumption (12.6%), decreased soda consumption (9.3%), and decreased likelihood of overweight or obese BMI levels (9.0%).

In contrast, within predominantly Hispanic zip codes we found a significant effect of higher, above-median, income on higher F&V consumption (5.7%), but not across the remaining three outcome variables. Higher education zip codes had the most positive association with diet health across all variables, with the exception of soda consumption for which none of the factors had a significant impact. Specifically, higher education was associated with increased F&V consumption (8.9%), decreased fast food consumption (11.9%), and decreased likelihood of overweight or obese BMI levels (13.7%). Higher grocery access and decreased fast food access had similar effects as on the overall population for some outcome variables (i.e. similar associations with likelihood of overweight or obese BMI levels and fast food consumption). However, in some cases the magnitude of impact was higher (i.e. grocery access increased Hispanic F&V consumption by 7.4% which more than twice the increase of the overall population) and in others, unlike the overall population, there was no significant association (i.e. no significant relationship between fast food/grocery access on soda consumption, or between fast food access and F&V consumption).

Few factors led to consistent improvements across all three subpopulations. Across all three groups, F&V consumption was strongly increased by high grocery access and education. Fast food consumption was reduced by all potential intervention targets besides increased income. Soda consumption was most reduced by decreased fast food access for Black and White-majority zip codes, but was not impacted by any of the intervention targets within Hispanic zip codes. Lastly, overweight or obese BMI levels were, by far, most strongly reduced across all groups by increased education levels.

4 Discussion

By analyzing 1.5 billion food intake logs and BMIs from 751,493 MyFitnessPal smartphone app users over seven years across 9,822 zip codes in relation to their demographic (education, ethnicity), socioeconomic (income), and food environment factors (grocery store and fast food access), we found 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. This is consistent with previous studies[36–38]. As suggested in Larson's review of neighborhood environment disparities in access to healthy foods in the U.S., residents who have better access to supermarkets and limited access to fast food restaurants have healthier diet and lower levels of obesity, and low-income, minority and rural neighborhoods often have limited access to supermarkets and higher availability of fast food restaurants[36]. Darmon[39] and Hiza[40] also found higher socioeconomic status (SES) groups had higher diet quality.

While we report on differences between above/below median, we find significantly larger effect sizes for top vs bottom quartile (Supplementary Figure e1). This indicates that effects of food environment on diet are even more pronounced when considering larger differences in income, education, access, and fast food prevalence. Of note, we found high education zip codes had the largest relative decrease in BMI levels of overweight and obese by 13.1%, which suggests that intervention programs and policies that aimed at increase education levels of residents of low-income neighborhoods are essential in curbing the obesity trends.

When we restricted our analyses to predominantly Black and Hispanic zip codes, we found the independent impacts of food access, income and education on food consumption and weight status varied significantly across Black, Hispanic and White populations. These findings suggest that tailored intervention strategies are needed based on neighborhood population distributions, assets and contexts.

Within predominantly Black zip codes, unlike other races, having higher income seems to be harmful as higher income Black zip codes had decreased fruits and vegetable consumption, increased fast food and soda consumption, and increased likelihood of overweight and obesity. The effect of higher education was low compared to grocery and fast food access. This could be explained by the "diminishing return hypothesis", which suggests that Blacks receive fewer protective health benefits from increases in SES than Whites[41,42]. Research has documented that Blacks and Whites receive different levels of economic return for their location in the educational and occupational hierarchies[43], and increases in SES (higher education and income) do not protect Blacks from increases in BMI in the same way as Whites[44]. A combination of factors, including neighborhood economic disadvantage[45,46], racial discrimination[47,48], and stress associated with educational attainment and mobility[49], may prevent higher SES Blacks from achieving their fullest health potential relative to Whites[44].

Furthermore, in Sherman's investigation of African American men's perceptions of their food environment, the author showed that African American men perceived that fast food was their food choice and that they had no other healthy food options in or near their residences. This perception was shaped by their food environment, food marketing and advertising, the cost of eating healthy versus convenience, as well as their formative years[23]. In fact, it has been documented that African American and Hispanic children were disproportionately exposed to more fast food advertisements than White children[50]. Grier and Kumanyikas study showed that Black youth were exposed to as much as 70% more fast food and soda TV commercials than White youth. Black youth also had higher media use and spending patterns, which made them ideal target for marketers[51]. While there have been many public nutrition education campaigns and individual level interventions designed to change individual dietary behavior, these cannot be matched for the amount of advertising and marketing of unhealthy food[52].

Within predominantly Hispanic zip codes, higher income was not associated with reduce BMI levels, but higher education significantly reduced BMI levels by 7%. High grocery access was the only factor associated with higher consumption of fresh fruits and vegetables, and high grocery and low fast food access had similar effects in reducing fast food and soda consumption and lowering overweight and obese BMI levels. The relationship between higher income and BMI could be partially explained by the "Hispanic health paradox" and "Hispanic health advantage"[53].

The Hispanic health paradox suggests that even though the first-generation Hispanics have lower SES, they experience better health outcomes including lower prevalence of cardiovascular diseases, asthma, diabetes and cancer compared to those who were U.S.-born[53–55]. Hispanic health advantage suggests that Hispanics have lower rates of harmful health behaviors, such as smoking, which in turn positively influence other health outcomes compared to non-Hispanic Whites[53,56–58]. Acculturation may also be another factor that influence Hispanics' dietary behaviors. Previous research has shown that by adopting American culture, Hispanic immigrants engage in less healthy behaviors, which in turn put themselves at higher risk for chronic diseases[53–55,59–63]. Obtaining higher education may increase the health literacy of Hispanics, which in turn could also influence their dietary behaviors, as well as preventive care use[64]. However, a recent study that examined atherosclerotic cardiovascular disease risk of a cohort of highly educated Hispanics found no evidence of the "Hispanic health paradox" in that Hispanics had the similar cardiovascular disease risk as non-Hispanic Whites with the same educational attainment, and this was likely due to acculturation[61].

Besides the individual health consequences, including increased risk of mortality, cardiovascular diseases, diabetes and certain cancers, obesity also brings direct and indirect economic consequences[65]. Cawley and Meyerhoefer estimated that obesity accounts for 21% of medical spending ($190 billion) in 2005[66], and if obesity trends continue to increase, obesity-related medical costs could rise by $48-66 billion a year[67]. More recent data indicated that the aggregate costs of obesity in the noninstitutionalized adult population of the US was as high as $315.8 billion in 2010, and the estimated extra annual medical care cost of an obese adult was $3,429 on average in 2013. The extra medical care costs are even higher for non-White populations ($4,086)[68]. Waters and Graf estimated chronic diseases caused obesity and overweight accounted for $480.7 billion in direct health care costs, plus additional $1.24 trillion in indirect costs attributed by lost productivity, which was 9.3 percent of the U.S. gross domestic product (GDP)[69].

The prevalence of overweight and obesity in the U.S., which was estimated to be 42.4% and about 107.6 million adults in 2017-2018[70]. A 13.1% decrease by implementing effective education programs and policies (i.e., based on our estimate of above/below median effect size, see Figure 4), could potentially lead to more than $48.3 billions of annual health care cost savings[71]. According to the U.S. Department of Education, the 2019 presidents education budget for the entire U.S.

was \$64 billion[72], which is significantly less than the aggregate costs of obesity. The national postsecondary education budget, which include important funding supporting upward mobility such as the federal Pell grant, work study and student loans, was \$24.2 billion in 2020[72]. Based on our analysis, implementing effective education programs and policies could potentially lead to more than \$48.3 billions of annual medical care cost savings in 2013 dollars. Taking inflation into account, this cost saving would be \$53.6 billion in 2020, and could support 83.8% of the education budget and cover the entire postsecondary budget twice. Effective program and policy interventions could range from mandatory schooling policies, such as those promoting health and nutrition education at schools, to increasing educational quality, such as those aimed at promoting higher levels of education[3,73,74].

Similarly, having higher grocery access, and lower fast food access could potentially lead to \$9.9 billions and \$6.1 billions of annual health care cost savings today respectively (based on our estimates in Figure 4). While it is challenging to close the education and income gaps, establishing more grocery stores and limit fast food restaurants together could potentially save \$16 billion. Previous reviews suggested that government policies that addressing food affordability and purchase, such as the Healthy Food Financing Initiative (HFFI), increasing food stamp (SNAP) benefit and provide incentives to create healthy retail food environment have been effective in reducing food insecurity and dietary behaviors[75–80]. While several studies showed that the establishments of new supermarkets had little improvement in BMI[81–83]; however, the investments in the new supermarkets have improved economic opportunity and social cohesion[84–86]. Our results showed that the impact of having higher grocery store access increased fresh fruit and vegetable consumption and decreased fast food consumption 2-3 times than for Whites. Although previous literature has shown null effects of grocery access[87,88], these studies have focused on the general population, which is White-skewed. Therefore, policies and strategies in increasing grocery store access and decreasing fast food access could potentially be the most effective approaches in changing dietary habits among African Americans.

Furthermore, having more grocery access and lower fast food access, in the food environment may work in synergistic ways that may lead to even lower obesity prevalence and obesity-related cost savings. This is demonstrated in a new study by Cantor et al. that HFFI boosted the effects of SNAP participation on improving food security and healthy food choices in food desserts[89].

This synergy could be multiplied when combining with effective education programs that could potentially lower obesity prevalence further by increasing individuals' SES (e.g., income and education)[90,91], health literacy and behaviors[90–94], as well as sense of control and empowerment[95].

Due to the cross-sectional design of this observational study, we were not able to make any causal inferences between SES and food environment variables and dietary behavior and BMI. However, we used a matching-based approach to mimic a quasi-experimental design to disentangle the individual impact of income, education and food access on participants' food consumption. Our analysis did not include other important variables such as gender and age, but we jointly considered the impacts of income, education and food environment access (grocery stores and fast food restaurants) on participants' food consumption with consistent measures across the U.S., whereas many previously published studies examined one or a few at a time. We used individuals' food loggings to estimate their consumption. Food loggings may not capture what individuals actually ate. However, we conducted rigorous validations through comparisons with high quality and highly representative datasets which demonstrated high correlations to gold-standard approaches (Figure 3). Majority of the food environment studies used screeners, food frequency questionnaires or 24-hour recalls for dietary assessment, and very few used diaries[7]. Our participants logged their food intakes for an average of 197 days each.

Our analysis was conducted at the zip code level. We used comparisons between above median and below-median zip codes across all dependent variables (SES and food environment) rather than using more specific cut-offs. In terms of zip code level income, our less wealthy zip codes are still relatively wealthy compared with average national income levels (Supplementary Table 4). Despite the fine-grained spatial resolution, the analysis included a total of 9,822 zip codes, covering the vast majority of US counties. Two large neighborhood food environment studies included 20,897 participants from the Reasons for Geographic and Racial Differences in Stroke Study[96] and 3,768 participants from the Framingham Heart Study[26] respectively. Recently, Aiello et al. conducted an analysis using 1.6 billion food purchases and 1.1 billion medical prescriptions for the city of London; however, as the authors point out, food purchase data could not be used to construct individual dietary patterns[97]. We used individual's food logging and BMI data generated from 1,164,926 participants using one of the most popular commercially available apps. Our study constitutes the largest United States nationwide study examining the impact of food environment on diet to date.

For this analysis, we also harnessed other large datasets such as Yelp to examine participants' food environments. The usage of our combined dataset goes beyond food environment and diet studies. Other attributes such as participants physical activity levels and social networks can be utilized to study other built and social environment exposures and health outcomes. participants location tracking data not only can be used to derive environmental exposures such as air pollution and noise[98], but also formulate personalized real-time intervention strategies[99,100]. Considering both the strengths and limitations of this study, more research is needed especially based on longitudinal study design and detailed individual level data to allow causal inference and precise interpretation of the results.

5 Conclusion

We analyzed 2.3 billion food intake logs and BMIs from 1.2 million MyFitnessPal smartphone app participants over seven years across 9,822 zip codes in relation to education, race/ethnicity, income, and food environment access. Our analyses indicated that 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 likelihood of being overweight or obese. Policy targeted at improving food access, income and education may increase healthy eating. Potential interventions may need to be targeted to specific subpopulations for optimal effectiveness.

ARTICLE INFORMATION

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

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.

Funding/Support: T.A. and J.L. were supported by a National Institutes of Health (NIH) grant (U54 EB020405, Mobilize Center, NIH Big Data to Knowledge Center of Excellence). T.A. was supported by the SAP Stanford Graduate Fellowship, NSF grant IIS-1901386, and Bill Melinda Gates Foundation (INV-004841). H.N was supported by NSF REU #1659585 at the Stanford Center for the Study of Language and Information (CSLI). J.H is supported by Postdoctoral Fellowship in Cardiovascular Disease Prevention (T32) funded by the National Heart, Lung, and Blood Institute (NHLBI) at the National Institutes of Health (NIH). J.L. was supported by DARPA under Nos. FA865018C7880 (ASED), N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions); Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Chan Zuckerberg Biohub, Amazon, Boeing, Chase, Docomo, Hitachi, Huawei, JD.com, NVIDIA, Dell.

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.

Additional Contributions: We thank MyFitnessPal for donating the data for independent research and Mitchell Gordon for helping with data preparation.

Bibliography

1. Yach, D., Hawkes, C., Gould, C. L. & Hofman, K. J. The global burden of chronic diseases: overcoming impediments to prevention and control. *JAMA* **291**, 2616–2622 (2004).

2. Beaglehole, R. *et al.* Priority actions for the non-communicable disease crisis. *The Lancet* **377**, 1438–1447 (2011).

3. Fleischer, N. L., Roux, A. V. D., Alazraqui, M. & Spinelli, H. Social patterning of chronic disease risk factors in a latin american city. *Journal of Urban Health* **85**, 923 (2008).

4. Sallis, J. F. & Glanz, K. The role of built environments in physical activity, eating, and obesity in childhood. *The future of children* **16**, 89–108 (2006).

5. Fleischhacker, S. E., Evenson, K. R., Rodriguez, D. A. & Ammerman, A. S. A systematic review of fast food access studies. *Obesity Reviews* **12**, e460–e471 (2011).

6. Odoms-Young, A., Singleton, C. R., Springfield, S., McNabb, L. & Thompson, T. Retail environments as a venue for obesity prevention. *Current obesity reports* **5**, 184–191 (2016).

7. Kirkpatrick, S. I. *et al.* Dietary assessment in food environment research: a systematic review. *American Journal of Preventive Medicine* **46**, 94–102 (2014).

8. Caspi, C. E., Sorensen, G., Subramanian, S. & Kawachi, I. The local food environment and diet: a systematic review. *Health & place* **18**, 1172–1187 (2012).

9. Feng, J., Glass, T. A., Curriero, F. C., Stewart, W. F. & Schwartz, B. S. The built environment and obesity: a systematic review of the epidemiologic evidence. *Health & place* **16**, 175–190 (2010).

10. Cummins, S. & Macintyre, S. Food environments and obesity—neighbourhood or nation? *International journal of epidemiology* **35**, 100–104 (2006).

11. Charreire, H. *et al.* Measuring the food environment using geographical information systems: a methodological review. *Public health nutrition* **13**, 1773–1785 (2010).

12. Kelly, B., Flood, V. M. & Yeatman, H. Measuring local food environments: an overview of available methods and measures. *Health & place* **17**, 1284–1293 (2011).

13. McKinnon, R. A., Reedy, J., Morrissette, M. A., Lytle, L. A. & Yaroch, A. L. Measures of the food environment: a compilation of the literature, 1990–2007. *American Journal of Preventive Medicine* **36**, S124–S133 (2009).

14. Elinder, L. S. & Jansson, M. Obesogenic environments–aspects on measurement and indicators. *Public Health Nutrition* **12**, 307–315 (2009).

15. Gittelsohn, J. & Sharma, S. Physical, consumer, and social aspects of measuring the food environment among diverse low-income populations. *American Journal of Preventive Medicine* **36**, S161–S165 (2009).

16. Gustafson, A., Hankins, S. & Jilcott, S. Measures of the consumer food store environment: a systematic review of the evidence 2000–2011. *Journal of community health* **37**, 897–911 (2012).

17. Lytle, L. A. Measuring the food environment: state of the science. *American Journal of Preventive Medicine* **36**, S134–S144 (2009).

18. Odoms-Young, A. M., Zenk, S. & Mason, M. Measuring food availability and access in african-american communities: implications for intervention and policy. *American Journal of Preventive Medicine* **36**, S145–S150 (2009).

19. Ohri-Vachaspati, P. & Leviton, L. C. Measuring food environments: a guide to available instruments. *American Journal of Health Promotion* **24**, 410–426 (2010).

20. Sharkey, J. R. Measuring potential access to food stores and food-service places in rural areas in the us. *American Journal of Preventive Medicine* **36**, S151–S155 (2009).

21. Kamphuis, C. B. *et al.* Environmental determinants of fruit and vegetable consumption among adults: a systematic review. *British Journal of Nutrition* **96**, 620–635 (2006).

22. Lytle, L. A. & Sokol, R. L. Measures of the food environment: a systematic review of the field, 2007–2015. *Health & Place* **44**, 18–34 (2017).

23. Kumanyika, S. K. Environmental influences on childhood obesity: ethnic and cultural influences in context. *Physiology & behavior* **94**, 61–70 (2008).

24. Sasson, C. *et al.* American heart association diabetes and cardiometabolic health summit: Summary and recommendations. *Journal of the American Heart Association* **7**, e009271 (2018).

25. Mason, K. E., Pearce, N. & Cummins, S. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from uk biobank. *The Lancet Public Health* **3**, e24–e33 (2018).

26. James, P., Seward, M. W., O'Malley, A. J., Subramanian, S. & Block, J. P. Changes in the food environment over time: examining 40 years of data in the framingham heart study. *International Journal of Behavioral Nutrition and Physical Activity* **14**, 84 (2017).

27. Bureau, U. C. American community survey 5-year estimates (2014). URL `http://censusreporter.org`. Accessed August 30, 2017.

28. Buajitti, E., Chiodo, S. & Rosella, L. C. Agreement between area-and individual-level income measures in a population-based cohort: Implications for population health research. *SSM-population health* **10**, 100553 (2020).

29. Rhone, A. Food access research atlas documentation (2015). URL `https://www.ers.usda.gov/data-products/food-access-research-atlas/documentation/`. Accessed August 20, 2017.

30. CDC. Behavioral risk factor surveillance system survey questionnaire (2011). URL `https://www.cdc.gov/brfss/smart/smart_2011.htm`. Accessed August 18, 2017.

31. CDC. Behavioral risk factor surveillance system survey questionnaire (2012). URL `https://www.cdc.gov/brfss/smart/smart_2012.htm`. Accessed August 18, 2017.

32. Einav, L., Leibtag, E. S. & Nevo, A. On the accuracy of nielsen homescan data. Tech. Rep. (2008).

33. Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* **46**, 399–424 (2011).

34. Stuart, E. A. Matching methods for causal inference: A review and a look forward. *Statistical Science: A review journal of the Institute of Mathematical Statistics* **25**, 1 (2010).

35. USCB. Quickfacts (2017). URL `https://www.census.gov/quickfacts/fact/table/US/PST045219`.

36. Larson, N. I., Story, M. T. & Nelson, M. C. Neighborhood environments: disparities in access to healthy foods in the us. *American Journal of Preventive Medicine* **36**, 74–81 (2009).

37. Mackenbach, J. D. *et al.* A systematic review on socioeconomic differences in the association between the food environment and dietary behaviors. *Nutrients* **11**, 2215 (2019).

38. Meyer, K. A. *et al.* Sociodemographic differences in fast food price sensitivity. *JAMA internal medicine* **174**, 434–442 (2014).

39. Darmon, N. & Drewnowski, A. Does social class predict diet quality? *The American journal of clinical nutrition* **87**, 1107–1117 (2008).

40. Hiza, H. A., Casavale, K. O., Guenther, P. M. & Davis, C. A. Diet quality of americans differs by age, sex, race/ethnicity, income, and education level. *Journal of the Academy of Nutrition and Dietetics* **113**, 297–306 (2013).

41. Farmer, M. M. & Ferraro, K. F. Are racial disparities in health conditional on socioeconomic status? *Social Science & Medicine* **60**, 191–204 (2005).

42. Shuey, K. M. & Willson, A. E. Cumulative disadvantage and black-white disparities in life-course health trajectories. *Research on Aging* **30**, 200–225 (2008).

43. Willson, A. E., Shuey, K. M. & Elder, G. H., Jr. Cumulative advantage processes as mechanisms of inequality in life course health. *American Journal of Sociology* **112**, 1886–1924 (2007).

44. Boen, C. The role of socioeconomic factors in black-white health inequities across the life course: Point-in-time measures, long-term exposures, and differential health returns. *Social Science & Medicine* **170**, 63–76 (2016).

45. Do, D. P. *et al.* Does place explain racial health disparities? quantifying the contribution of residential context to the black/white health gap in the united states. *Social Science & Medicine* **67**, 1258–1268 (2008).

46. Williams, D. R. & Collins, C. Us socioeconomic and racial differences in health: patterns and explanations. *Annual Review of Sociology* **21**, 349–386 (1995).

47. Bratter, J. L. & Gorman, B. K. Is discrimination an equal opportunity risk? racial experiences, socioeconomic status, and health status among black and white adults. *Journal of Health and Social Behavior* **52**, 365–382 (2011).

48. Williams, D. R., Neighbors, H. W. & Jackson, J. S. Racial/ethnic discrimination and health: findings from community studies. *American Journal of Public Health* **93**, 200–208 (2003).

49. Pearson, J. A. Can't buy me whiteness: New lessons from the titanic on race, ethnicity, and health. *Du Bois Review: Social Science Research on Race* **5**, 27–47 (2008).

50. Powell, L. M., Zhao, Z. & Wang, Y. Food prices and fruit and vegetable consumption among young american adults. *Health & place* **15**, 1064–1070 (2009).

51. Grier, S. A. & Kumanyika, S. K. The context for choice: health implications of targeted food and beverage marketing to african americans. *American journal of public health* **98**, 1616–1629 (2008).

52. Crawford, P. B. Nutrition policies designed to change the food environment to improve diet and health of the population. In *Nutrition Education: Strategies for Improving Nutrition and Healthy Eating in Individuals and Communities*, vol. 92, 107–117 (Karger Publishers, 2020).

53. Lommel, L. L., Thompson, L., Chen, J.-L., Waters, C. & Carrico, A. Acculturation, inflammation, and self-rated health in mexican american immigrants. *Journal of Immigrant and Minority Health* **21**, 1052–1060 (2019).

54. Bostean, G. Does selective migration explain the hispanic paradox? a comparative analysis of mexicans in the us and mexico. *Journal of Immigrant and Minority Health* **15**, 624–635 (2013).

55. Ruiz, J. M., Steffen, P. & Smith, T. B. Hispanic mortality paradox: a systematic review and meta-analysis of the longitudinal literature. *American Journal of Public Health* **103**, e52–e60 (2013).

56. Min, J. W., Rhee, S., Lee, S. E., Rhee, J. & Tran, T. Comparative analysis on determinants of self-rated health among non-hispanic white, hispanic, and asian american older adults. *Journal of Immigrant and Minority Health* **16**, 365–372 (2014).

57. Kimbro, R. T., Gorman, B. K. & Schachter, A. Acculturation and self-rated health among latino and asian immigrants to the united states. *Social Problems* **59**, 341–363 (2012).

58. Brewer, J. V. *et al.* Contributors to self-reported health in a racially and ethnically diverse population: focus on hispanics. *Annals of Epidemiology* **23**, 19–24 (2013).

59. Ridker, P. M. C-reactive protein: a simple test to help predict risk of heart attack and stroke. *Circulation* **108**, e81–e85 (2003).

60. Viruell-Fuentes, E. A. Beyond acculturation: immigration, discrimination, and health research among mexicans in the united states. *Social Science & Medicine* **65**, 1524–1535 (2007).

61. Rodriguez, F. *et al.* Association of educational attainment and cardiovascular risk in hispanic individuals: Findings from the cooper center longitudinal study. *JAMA Cardiology* **4**, 43–50 (2019).

62. Lutsey, P. L. *et al.* Associations of acculturation and socioeconomic status with subclinical cardiovascular disease in the multi-ethnic study of atherosclerosis. *American Journal of Public Health* **98**, 1963–1970 (2008).

63. Koya, D. L. & Egede, L. E. Association between length of residence and cardiovascular disease risk factors among an ethnically diverse group of united states immigrants. *Journal of General Internal Medicine* **22**, 841–846 (2007).

64. Daviglus, M. L., Pirzada, A. & Talavera, G. A. Cardiovascular disease risk factors in the hispanic/latino population: lessons from the hispanic community health study/study of latinos (hchs/sol). *Progress in Cardiovascular Diseases* **57**, 230–236 (2014).

65. CDC. Adult obesity causes and consequences (2017). URL `https://www.cdc.gov/obesity/adult/causes.html`.

66. Cawley, J. & Meyerhoefer, C. The medical care costs of obesity: an instrumental variables approach. *Journal of Health Economics* **31**, 219–230 (2012).

67. Wang, Y. C., McPherson, K., Marsh, T., Gortmaker, S. L. & Brown, M. Health and economic burden of the projected obesity trends in the usa and the uk. *The Lancet* **378**, 815–825 (2011).

68. Biener, A., Cawley, J. & Meyerhoefer, C. The high and rising costs of obesity to the us health care system (2017).

69. Waters, H. & Graf, M. America's obesity crisis: The health and economic costs of excess weight. Available at `https://www.milkeninstitute.org/reports/americas-obesity-crisis-health-and-economic-costs-excess-weight`, note=Accessed March 05, 2020,.

70. CDC. Prevalence of obesity and severe obesity among adults: United states, 2017–2018 (2018). URL `https://www.cdc.gov/nchs/products/databriefs/db360.htm`.

71. Census Quick Facts: Population. `https://www.census.gov/quickfacts/fact/table/US/PST045218`. Accessed March 05, 2020.

72. of Education, U. D. Budget history tables (2020). URL `https://www2.ed.gov/about/overview/budget/tables.html?src=rt`.

73. Cohen, A. K., Rai, M., Rehkopf, D. H. & Abrams, B. Educational attainment and obesity: a systematic review. *Obesity Reviews* **14**, 989–1005 (2013).

74. Cohen, A. K. & Syme, S. L. Education: a missed opportunity for public health intervention. *American Journal of Public Health* **103**, 997–1001 (2013).

75. Moran, A. *et al.* Financial incentives increase purchases of fruit and vegetables among lower-income households with children. *Health Affairs* **38**, 1557–1566 (2019).

76. Moran, A. J. *et al.* Associations between governmental policies to improve the nutritional quality of supermarket purchases and individual, retailer, and community health outcomes: An integrative review. *International journal of environmental research and public health* **17**, 7493 (2020).

77. Phipps, E. J. *et al.* Impact of a rewards-based incentive program on promoting fruit and vegetable purchases. *American journal of public health* **105**, 166–172 (2015).

78. Steele-Adjognon, M. & Weatherspoon, D. Double up food bucks program effects on snap recipients' fruit and vegetable purchases. *BMC public health* **17**, 1–7 (2017).

79. Wilde, P., Klerman, J. A., Olsho, L. E. & Bartlett, S. Explaining the impact of usda's healthy incentives pilot on different spending outcomes. *Applied Economic Perspectives and Policy* **38**, 655–672 (2015).

80. Polacsek, M. *et al.* A supermarket double-dollar incentive program increases purchases of fresh fruits and vegetables among low-income families with children: the healthy double study. *Journal of nutrition education and behavior* **50**, 217–228 (2018).

81. Cummins, S., Flint, E. & Matthews, S. A. New neighborhood grocery store increased awareness of food access but did not alter dietary habits or obesity. *Health affairs* **33**, 283–291 (2014).

82. Dubowitz, T. *et al.* Changes in diet after introduction of a full service supermarket in a food desert. *Health Affairs (Project Hope)* **34**, 1858 (2015).

83. Zhang, Y. T. *et al.* Is a reduction in distance to nearest supermarket associated with bmi change among type 2 diabetes patients? *Health & place* **40**, 15–20 (2016).

84. Rogus, S., Athens, J., Cantor, J. & Elbel, B. Measuring micro-level effects of a new supermarket: do residents within 0.5 mile have improved dietary behaviors? *Journal of the Academy of Nutrition and Dietetics* **118**, 1037–1046 (2018).

85. Chrisinger, B. A mixed-method assessment of a new supermarket in a food desert: contributions to everyday life and health. *Journal of Urban Health* **93**, 425–437 (2016).

86. Giang, T., Karpyn, A., Laurison, H. B., Hillier, A. & Perry, R. D. Closing the grocery gap in underserved communities: the creation of the pennsylvania fresh food financing initiative. *Journal of Public Health Management and Practice* **14**, 272–279 (2008).

87. Jack, D. *et al.* Socio-economic status, neighbourhood food environments and consumption of fruits and vegetables in new york city. *Public health nutrition* **16**, 1197–1205 (2013).

88. Strome, S., Johns, T., Scicchitano, M. J. & Shelnutt, K. Elements of access: the effects of food outlet proximity, transportation, and realized access on fresh fruit and vegetable consumption in food deserts. *International quarterly of community health education* **37**, 61–70 (2016).

89. Cantor, J. *et al.* Snap participants improved food security and diet after a full-service supermarket opened in an urban food desert: Study examines impact grocery store opening had on food security and diet of supplemental nutrition assistance program participants living in an urban food desert. *Health Affairs* **39**, 1386–1394 (2020).

90. Chandola, T., Clarke, P., Morris, J. & Blane, D. Pathways between education and health: a causal modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* **169**, 337–359 (2006).

91. Cutler, D. M. & Lleras-Muney, A. Education and health: evaluating theories and evidence. Tech. Rep., National Bureau of Economic Research (2006).

92. Kenkel, D. S. Health behavior, health knowledge, and schooling. *Journal of Political Economy* **99**, 287–305 (1991).

93. Fletcher, J. M. & Frisvold, D. E. Higher education and health investments: does more schooling affect preventive health care use? *Journal of Human Capital* **3**, 144–176 (2009).

94. Lleras-Muney, A. The relationship between education and adult mortality in the united states. *The Review of Economic Studies* **72**, 189–221 (2005).

95. Mirowsky, J. *Education, Social Status, and Health* (Routledge, 2017).

96. Chen, M. *et al.* Association of community food environment and obesity among us adults: a geographical information system analysis. *J Epidemiol Community Health* **73**, 148–155 (2019).

97. Aiello, L. M., Schifanella, R., Quercia, D. & Del Prete, L. Large-scale and high-resolution analysis of food purchases and health outcomes. *EPJ Data Science* **8**, 14 (2019).

98. Khan, J., Ketzel, M., Kakosimos, K., Sørensen, M. & Jensen, S. S. Road traffic air and noise pollution exposure assessment–a review of tools and techniques. *Science of The Total Environment* **634**, 661–676 (2018).

99. Hardeman, W., Houghton, J., Lane, K., Jones, A. & Naughton, F. A systematic review of just-in-time adaptive interventions (jitais) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity* **16**, 31 (2019).

100. Kankanhalli, A., Saxena, M. & Wadhwa, B. Combined interventions for physical activity, sleep, and diet using smartphone apps: A scoping literature review. *International journal of medical informatics* **123**, 54–67 (2019).

101. Rahkovsky, I. & Snyder, S. Food choices and store proximity (2015). URL `https://www.ers.usda.gov/webdocs/publications/45432/53943_err195.pdf?v=42276/`. Accessed December 31, 2019.

102. Wikipedia. List of fast food restaurant chains — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=List%20of%20fast%20food%20restaurant%20chains&oldid=933158460` (2017). [Online; accessed 01-September-2017].

103. Wikipedia. List of restaurant chains in the United States — Wikipedia, the free encyclopedia. `http://en.wikipedia.org/w/index.php?title=List%20of%20restaurant%20chains%20in%20the%20United%20States&oldid=930612669` (2017). [Online; accessed 01-September-2017].

104. Hartlaub, P. Sweet! america's top 10 brands of soda. http://www.nbcnews.com/id/42255151/ns/business-us_business/t/sweet-americas-top-brands-soda (2017). [Online; accessed 01-September-2017].

105. CDC. Healthy weight, overweight, and obesity among u.s. adults (2018). URL https://www.cdc.gov/nchs/data/nhsr/nhsr122-508.pdf.

106. CDC. Obesity and overweight (2016). URL https://www.cdc.gov/nchs/fastats/obesity-overweight.htm.

107. Wilson, J. Americans getting older, new census figures show (2019). URL https://thehill.com/homenews/state-watch/449505-americans-getting-older-new-census-figures-show.

108. Worldbank. Population, female (% of total population) (2019). URL https://data.worldbank.org/indicator/SP.POP.TOTL.FE.ZS.

109. USCB. Income, poverty and health insurance coverage in the united states: 2016 (2017). URL https://www.census.gov/newsroom/press-releases/2017/income-povery.html.

110. Wilson, J. Census: More americans have college degrees than ever before (2017). URL https://thehill.com/homenews/state-watch/326995-census-more-americans-have-college-degrees-than-ever-before.

111. Din, A. Hud usps zip code crosswalk files (2017). URL https://www.huduser.gov/portal/datasets/usps_crosswalk.html. Accessed August 10, 2017.

112. Yelp. Yelp Fusion V3 API (2017). URL https://www.yelp.com/dataset/documentation/main/. Accessed August 30, 2017.

113. Efron, B. & Tibshirani, R. J. *An introduction to the bootstrap* (CRC press, 1994).

114. Austin, P. C. & Small, D. S. The use of bootstrapping when using propensity-score matching without replacement: a simulation study. *Statistics in medicine* **33**, 4306–4319 (2014).

115. Diamond, A. & Sekhon, J. S. Genetic matching for estimating causal effects (2012).

116. Teixeira, V., Voci, S. M., Mendes-Netto, R. S. & da Silva, D. G. The relative validity of a food record using the smartphone application myfitnesspal. *Nutrition & Dietetics* **75**, 219–225 (2018).

117. Griffiths, C., Harnack, L. & Pereira, M. A. Assessment of the accuracy of nutrient calculations of five popular nutrition tracking applications. *Public Health Nutrition* **21**, 1495–1502 (2018).

118. Chen, J., Berkman, W., Bardouh, M., Ng, C. Y. K. & Allman-Farinelli, M. The use of a food logging app in the naturalistic setting fails to provide accurate measurements of nutrients and poses usability challenges. *Nutrition* **57**, 208–216 (2019).

119. Nguyen, Q. *et al.* Social media indicators of the food environment and state health outcomes. *Public Health* **148**, 120–128 (2017).

120. Gomez-Lopez, I. N. *et al.* Using social media to identify sources of healthy food in urban neighborhoods. *Journal of Urban Health* **94**, 429–436 (2017).

121. Glass, G. V. Primary, secondary, and meta-analysis of research. *Educational Researcher* **5**, 3–8 (1976).

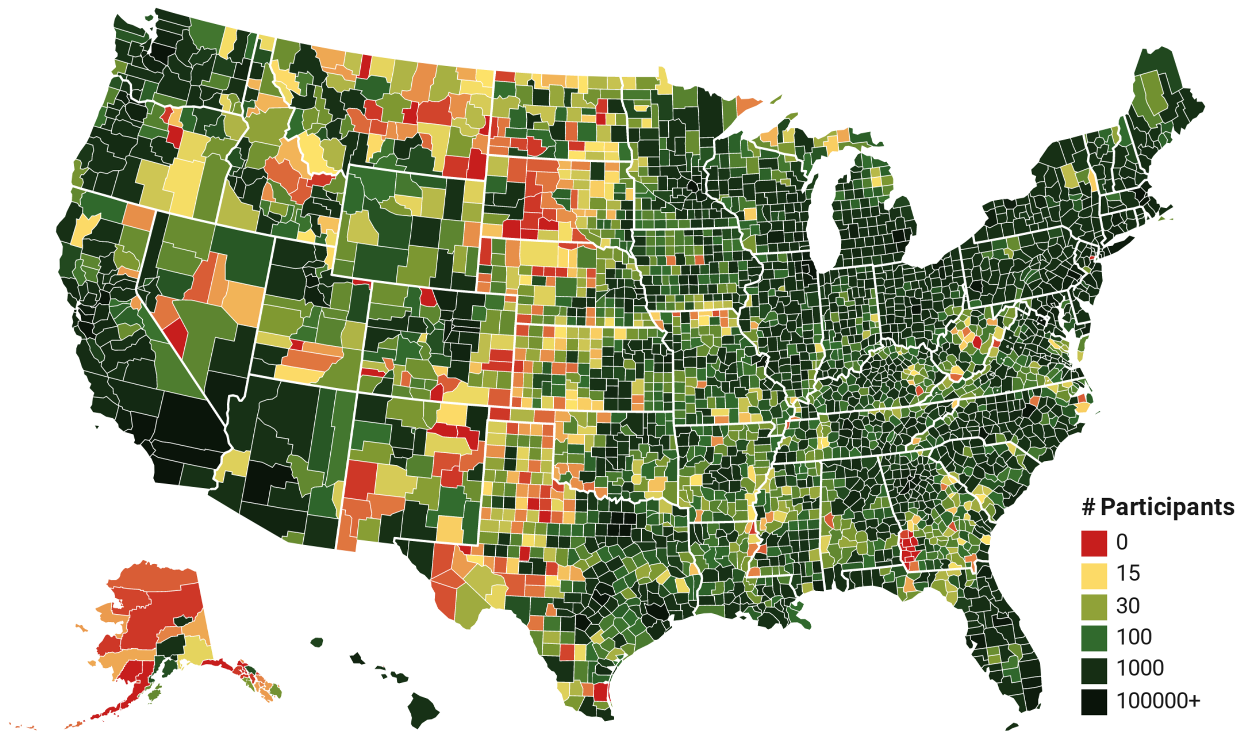

Figure 1: **Number of participants in our study across U.S. counties.** A choropleth showing the number of participants in each U.S. county. This country-wide observational study included 1,164,926 participants across 9,822 U.S. zip codes that collectively logged 2.3 billion food entries for an average of 197 days each. Most U.S. counties are represented by at least 30 participants in our dataset. This study constitutes the largest nationwide study examining the impact of food environment on diet to date, with 300 times more participants and 4 times more person years of tracking than the Framingham Heart Study[26].

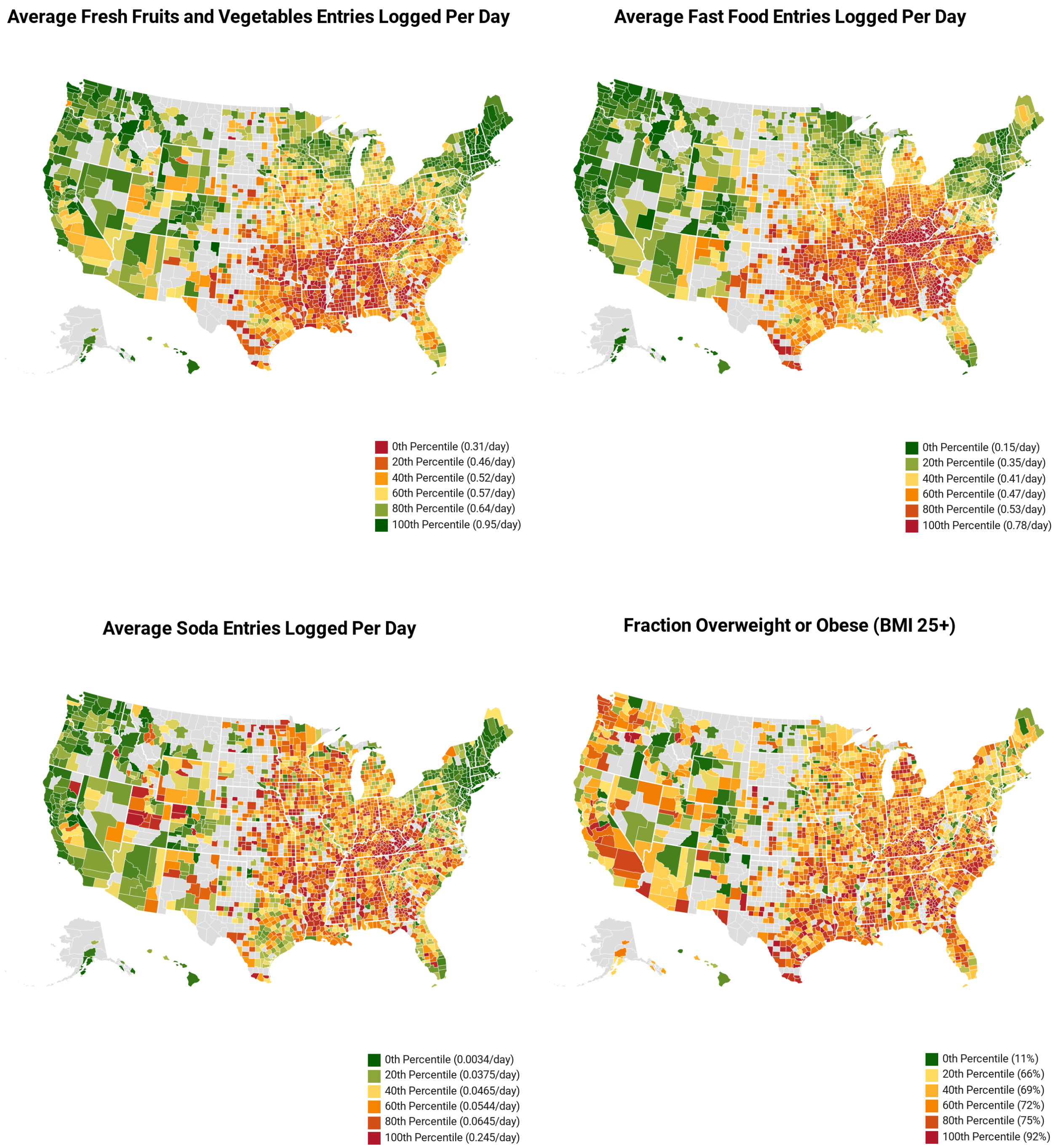

Figure 2: **Dietary Consumption and Body Weight across U.S. Counties.** A set of choropleths showing the main study outcomes of the number of entries that are classified as fresh fruit and vegetables, fast food, and soda consumption as well as the fraction of overweight/obese participants across the USA by counties with more than 30 participants. We observe that food consumption healthfulness varies significantly across counties in the United States.

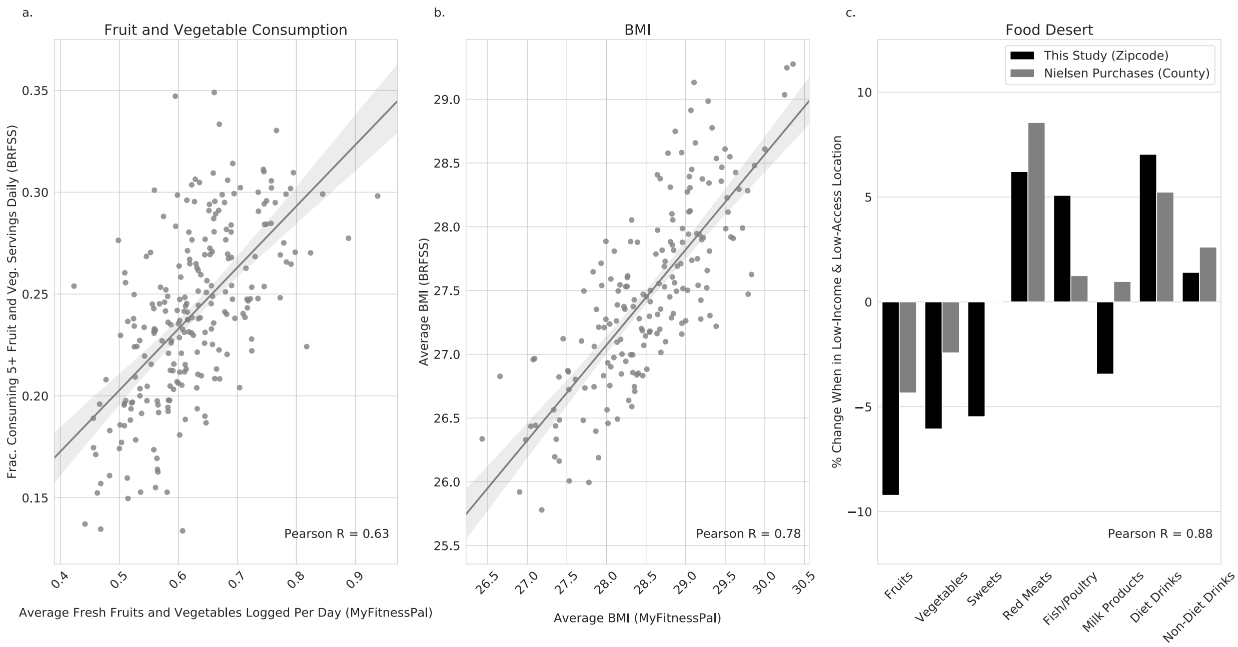

Figure 3: **This studies' smartphone-based food logs correlate with previous, representative survey measures and purchase data. a,** Fraction of fresh fruits and vegetables logged is correlated with BRFSS survey data[30] (R=0.63, $p < 10^{-5}$; Methods). **b,** Body Mass Index of smartphone cohort is correlated with BRFSS survey data[31] (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[101] (R=0.88, $p < 0.01$; Methods). These results demonstrate that smartphone-based food logs are highly correlated with existing, gold-standard survey measures and purchase data.

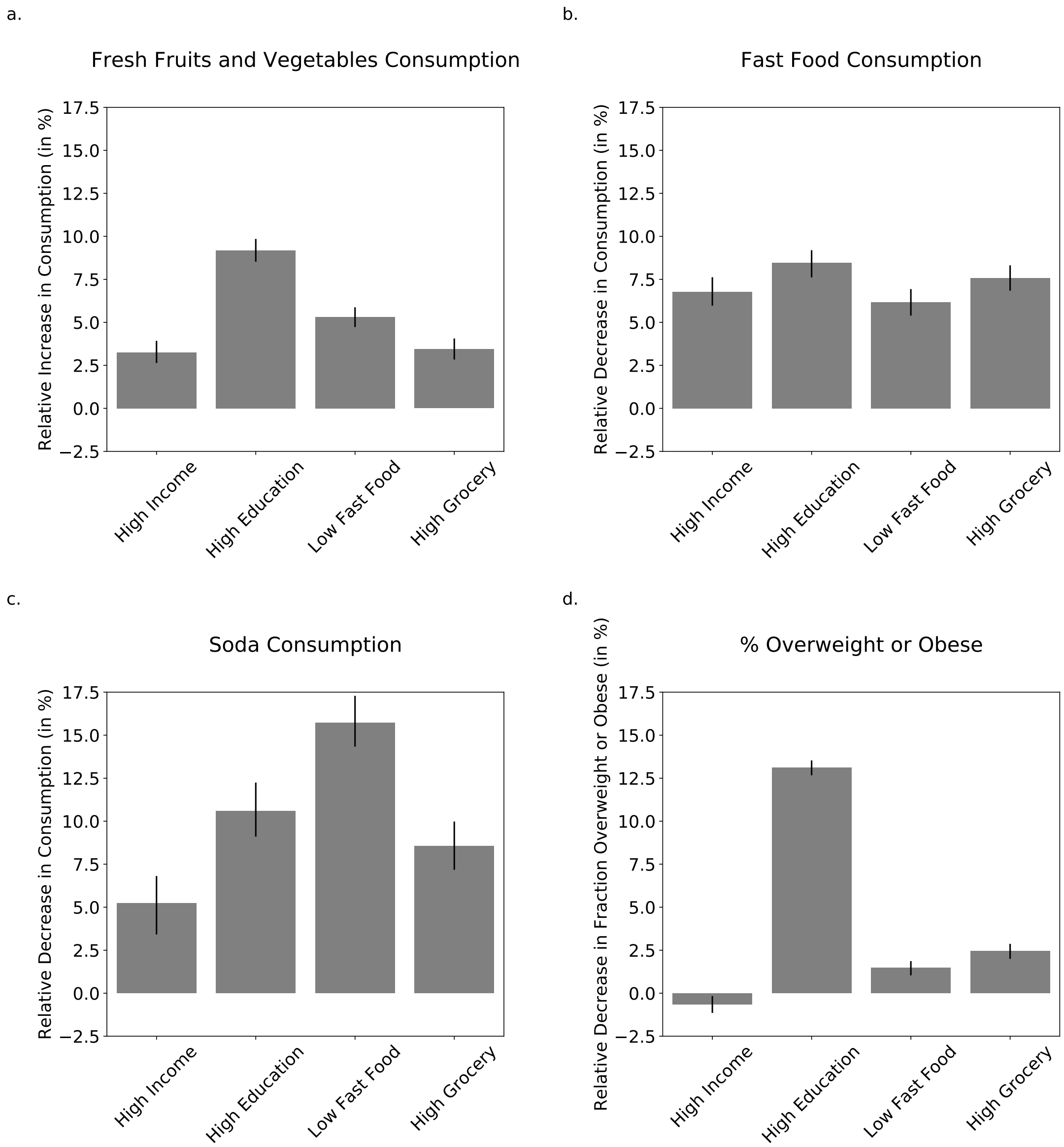

Figure 4: **The impact of income, education, grocery store and fast food access on food consumption and weight status.** Independent contributions of high income (median family income higher than or equal to \$70,241), higher education (fraction of population with college education 29.8% or higher), high grocery access (fraction of population that is closer than 0.5 miles from nearest grocery store is greater than or equal to than 20.3%), and low fast food access (less than or equal to 5.0% 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 percent overweight or obese (BMI 25+). Cut points correspond to median values. Estimates are based on matching experiments controlling for all but the one treatment variable in focus (Methods). Error bars correspond to bootstrapped 95% confidence intervals (Supplementary Methods). While the most impactful factors vary across outcomes, only higher education was associated with a sizeable decrease of 13.1% in overweight and obese weight status.

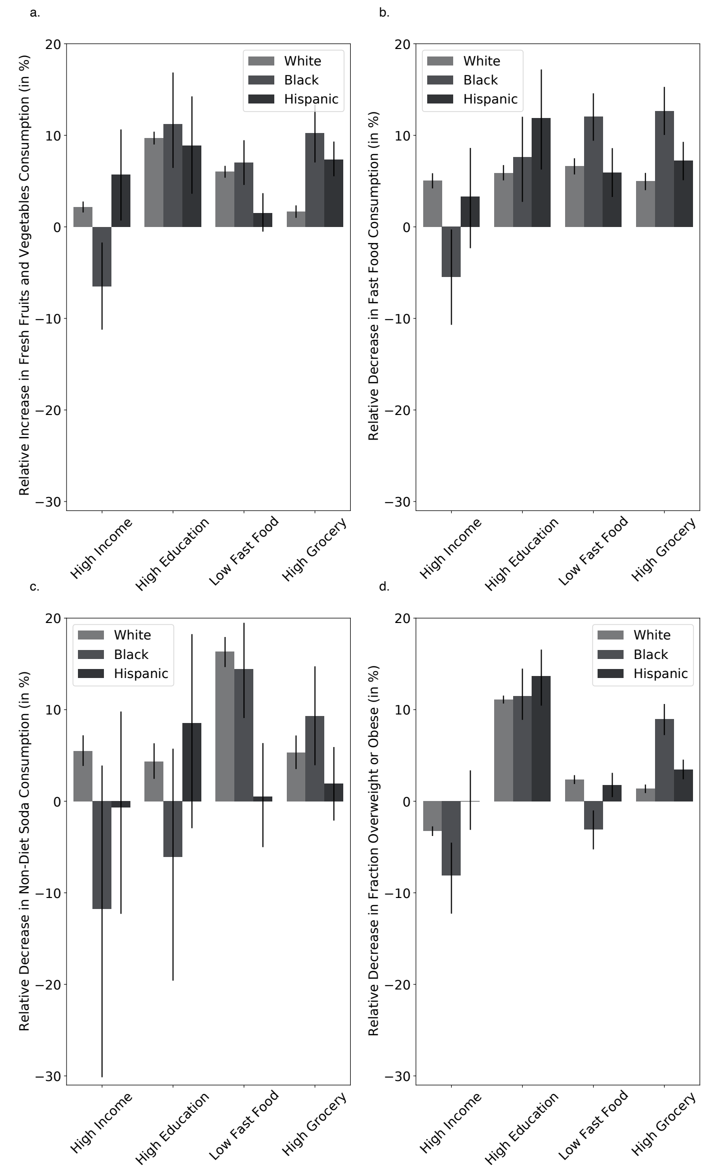

Figure 5: **Effects on food consumption and weight status disaggregated by predominantly Black, Hispanic, and non-Hispanic White zip codes (*i.e.*, 50% or more).** Independent contributions of high income (median family income higher than or equal to \$70,241), higher education (fraction of population with college education 29.8% or higher), high grocery access (fraction of population that is closer than 0.5 miles from nearest grocery store is greater than or equal to than 20.3%), and low fast food access (less than or equal to 5.0% 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 percent overweight or obese (BMI $>$ 25). Cut points correspond to median values. Estimates are based on matching experiments controlling for all but one treatment variable (Methods). Error bars correspond to bootstrapped 95% confidence intervals (Methods). We observe significant differences across Black, Hispanic, and non-Hispanic White zip codes in terms how food consumption is affected by factors of income, education, fast food and grocery access.

Supplemental Materials

- Supplemental Methods
- Supplemental Figures
- Supplemental Tables

List of Figures

List of Tables

Supplemental Methods

Details on 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[29]. A census tract is considered a food desert by the USDA if it is both low-income (defined by Department of Treasury's New Markets Tax Credit program) and low-access, meaning at least 500 people or 30 percent of residents live more than 0.5 miles from a supermarket in urban areas (10 miles in rural areas)[101].

Although the USDA uses different thresholds for urban and rural areas (0.5 miles and 10 miles respectively), we found that even in non-Urban zipcodes (defined as USDA rural-urban continuum RUCA scores of 4 or higher), the fraction of that population that is farther than 0.5 miles from grocery stores had the highest correlation to Fruit Vegetable consumption (R=-0.19), compared to 1 miles (R=-0.13), 10 miles (R=-0.07), and 20 miles (R=0.02). This suggests that the fraction of the population farther than 0.5 miles from a grocery store has the strongest relationship with healthy food consumption, *even in non-Urban zipcodes*. Hence, we decided used 0.5 miles distance as a standard measure of grocery access for both rural and urban zip codes, contrary to the USDA definition. We subsequently sanity checked for any downstream confounding of urbanicity in our primary matching experiment of above/below median grocery access, and found a negligible difference (Standardized Mean Difference of 0.18) in urbanicity between control and treatment, suggesting that the effect size was not due to grocery store distance functioning as a proxy for urbanicity, but rather directly due to differential grocery access.

We aggregated these data from a census tract level to a zip code level using USPS Crosswalk data, which provides a list of all census tracts which overlap with a single zip code[111]. We related these data on census tract level to the zip code level by taking the weighted average of each census

Table e1: Demographic statistics for our study compared with nationally representative survey data. (*) indicates statistics calculated at the zip code level.

Source	BMI	Overweight or Obese	Obese	Median Age	Gender	Median Family Income*	College*	Ethnicity*
Our Study	28.5	67.8%	32.8%	36	74% Female	$ 76,563	33.7%	68.3% White
Nat.l Avg.	29.4[105]	71.6%[106]	39.8%[106]	38.2[107]	50.5% Female[108]	$ 59,039[109]	33.4%[110]	61.3% White[35]

tract food environment measure (both grocery store access and food desert status), weighted by the number of people in the tract[111]. For instance, if zip code A overlapped with Census Tract A (2500 people, food desert) and Census Tract B (7500 people, not a food desert), the food desert measure of zip code A would be estimated as 25%. We defined the binary threshold for food desert, used in Figure 3, as 50% or higher.

We measured fast food access through the fraction of restaurants in a zip-code that are fast food restaurants. Data on local restaurants and businesses were obtained through the Yelp API[112]. 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). This resulted in a varying sample radius depending on urbanicity. For example, Urban zipcodes (RUCA code of 1) had an average effective centroid size of 15 miles, which we calculated by taking the distance from the zipcode center to the furthest restaurant returned by Yelp. We further used Yelp-based environment variables that we expected *not* to influence food consumption, such as the availability of waterproofing services, countertop installers, or electronic stores, as null experiments to demonstrate discriminant validity of our statistical analysis pipeline (see Supplementary Figure e2).

Details on outcome measure (food consumption and weight status) We used 2.3 billion food intake logs by 1,164,926 U.S. participants of the MyFitnessPal (MFP) smartphone application to quantify food consumption across 9,822 zip codes. During the observation period from January 1, 2010 to November 15, 2016, the average participant logged 9.30 entries into their digital food journal per day. The average participant used the app for 197 days. All participants in this sample used the app for at least 10 days. Each entry contains a separate food component (*e.g.*, banana, yogurt, hamburger, ...). We classified all entries into three categories: fresh fruits and vegetables (F&V; through a proprietary classifier by MFP), fast food (if the description contained the name of a fast food chain listed in Supplementary Information Table e6, and sugary (non-diet) soda (if the description contained the name of a soda drink listed in Supplementary Table e7 but did not contain “diet”, “lite”, “light”, or “zero”). In all cases, descriptions, as well fast food and soda drink keywords, were normalized by lower-casing and removing punctuation. We then calculated the average number of food entries logged per participant, per day, for each of the F&V, fast food, and soda categories (e.g. average number of F&V logged per participant per day), excluding days in

Table e2: Outcome measures calculated at the zip code level for the **9,822** zip codes in our study, spanning **1,164,926** participants.

	# Participants	F&V Entries per Day	Soda Entries per Day	Fast Food Entries per Day	BMI	% Overweight or Obese	% Obese
mean	118.6	0.61	0.04	0.39	28.8	69	35
median	90	0.60	0.04	0.38	28.8	70	35
std	91.5	0.11	0.02	0.10	1.6	10	11
min	30	0.25	0.001	0.12	23.2	17	2
max	1262	1.29	0.17	0.91	36.9	100	80

which the participant was inactive (*i.e.*, consistently did not log anything). Finally, we aggregated these participant-level measures to the zip code level by taking the mean of each category's measure for all participants in each zip code. We further used body-mass index (BMI) health in each zip code as a weight status outcome. We operationalize BMI health as the fraction of participants in a zip code which are overweight or obese (BMI > 25). BMI was self-reported by participants of the smartphone application (99.92% of participants did report BMI). Supplementary Table e2 shows basic summary statistics for the outcome measures used in this study. In our statistical analyses, we compared two sets of zip codes that differ in a dimension of interest (*e.g.*, grocery store access access) as treatment and control group and use the relative change in F&V consumption, fast food consumption, soda consumption, and BMI health of the treated group relative to the control group. To generate confidence intervals, as well as to compute p-values to test for statistical significance of changes in outcome, we use non-parametric bootstrap resampling[113]. Specifically, we follow the method proposed by Austin and Small,[114] which is to draw bootstrap samples post-matching from the matched pairs in the propensity-score-matched sample after the Genetic Matching stage[115]. We confirmed the validity of this method empirically by also calculating t-tests for each experiment, which gave qualitatively similar results.

Data Validation We find that our study population has significant overlap with the U.S. national population (Supplementary Table e1) but is skewed towards women and higher income. We demonstrate that food consumption measured based on this population are highly correlated with state-of-the-art measures (Figure 3, Section 2.3). 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 hour dietary recalls and food composition tables. These studies found that both measures tend to be highly correlated[116,117], but that app-reported measures tend to underestimate certain macro- and micronutrients[116,117], especially in populations that were previously unfamiliar with the smartphone applications[118]. In contrast, this study leverages a convenience sample of existing participants of the smartphone app MyFitnessPal. Yelp data has been used in measures of food environment[119] and a study in Detroit found Yelp data to be more accurate than commercially-available databases such as Reference USA[120]. 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 Mexican food consumption (entries labeled as Mexican food by a proprietary MFP classifier, logged per participant, per day) was correlated with the fraction of Mexican restaurants (R=0.72; $< 10^{-4}$) 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=0.51; $P < 10^{-4}$).

Details on reproducing state-of-the-art measures using population-scale digital food logs We used the latest survey data from BRFSS[30,31] available at the county-level. Specifically, we used variables FV5SRV from BRFSS 2011 representing the the faction of people eating five or more servings of fresh fruit and vegetables[30], and BMI5 from BRFSS 2012 representing body mass index[31]. We compare against BRFSS rather than National Health and Nutrition Examination Survey (NHANES), since BRFSS is significantly larger than NHANES, it is remotely administered matching our study, and it has much better geographical coverage than NHANES and geographical comparisons are central to our study. 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 (Figure 3a, R=0.63, $p < 10^{-5}$). Further, average county-level BMI was strongly correlated as well (Figure 3b, R=0.78, $p < 10^{-5}$). We further compared to published results by the USDA[101], which used data from the 2010 Nielsen Homescan Panel Survey that captured household food purchases for in-home consumption (but did not capture 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[101] across categories of fruit, vegetable, sweets, red meat, fish/poultry, milk products, diet drinks, and non-diet drinks (Table 4 in Rahkovsky and Snyder[101]). 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 (Figure 3c, R=0.88, $p < 0.01$) 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.

Details on Matching Approach Specifically, we use a one-to-one Genetic Matching approach[115], with replacement, and use the mean of the Standardized Mean Difference (SMD) between treatment and control groups, across all matched variables, as the Genetic Matching balance metric in order to maximize balance (overlap) between the treated and the control units. Some definitions of SMD use the standard variation in the overall population before matching[33]. 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[121].

After matching, we evaluated the quality of balance between the treated and the control units by the Standardized Mean Difference across each of the variables that were controlled for and included in the matching process. A good balance between treated and control groups was defined as a Standardized Mean Difference (SMD) of less than 0.25 standard deviations[34] across each variable. By default, we do not enforce a caliper in order to minimize bias in matching process, although in rare cases in which a good balance was not achieved, a caliper was enforced, starting at 2.5 standard deviations between matched and controlled units, and decreased by 0.1 until the matched and control groups had a SMD smaller than 0.25 across all matched variables.

For the vast majority of matching experiments the SMD across all matched variables was well below 0.25, with a mean of 0.040 and median of 0.016 for the four overall population matching experiments. The SMD for the race-majority zipcode experiments was slightly higher, but still very significantly below 0.25 across all 12 experiments, with a mean of 0.055 and median of 0.036.

Thus, no caliper was necessary to ensure a good balance, with the exception of one out of the 12 of sub-population experiments (White, High Education). Detailed balancing statistics for each of the matches are available in the Appendix (Tables e8a-e30b), as well as a supplementary matching experiment in which a top/bottom quartile split was used instead of a median split (Figure e1).

Details on obesity prevention related cost savings It was estimated that the aggregate costs of obesity in the non-institutionalized adult population of the U.S. was as high as $315.8 billion in 2010, and the estimated extra annual medical care cost of an obese adult was $3,429 on average in 2013. The extra medical care costs are even higher for non-White populations ($4,086)[68]. The prevalence of overweight and obesity among US adults was 42.4% and about 107.6 million adults in 2017-2018[70]. A 13.1% decrease by implementing effective education programs and policies (*i.e.*, based on our estimate of above/below median effect size, see Figure 4), could potentially lead to more than $48.3 billions of annual health care cost savings[71]. To produce this estimate we need to assume that the Average Treatment Effect on the Treated (ATT; estimated through our matching procedure) can be generalized to the overall U.S. population. According to the U.S. Department of Education, the 2019 president's education budget for the entire U.S. was $64 billion[72], which is significantly less than the aggregate costs of obesity. The national postsecondary education budget, which include important funding supporting upward mobility such as the federal Pell grant, work study and student loans, was $24.2 billion in 2020[72]. Based on our analysis, implementing effective education programs and policies could potentially lead to more than $48.3 billions of annual medical care cost savings in 2013 dollars. Taking inflation into account, this cost saving would be $53.6 billion in 2020, and could support 83.8% of the education budget and cover the entire postsecondary budget twice. Effective program and policy interventions could range from mandatory schooling policies, such as those promoting health and nutrition education at schools, to increasing educational quality, such as those aimed at promoting higher levels of education[3,73,74]. Similarly, having higher grocery access (2.4% decrease in overweight and obesity), and lower fast food access (1.5% decrease in overweight and obesity) could potentially lead to $9.9 billions and $6.1 billions of annual health care cost savings today respectively (based on our estimates in Figure 4).

Table e3: Effect sizes of all top/bottom half matching experiments (Fig. 4).

Treatment	Outcome	% Change	Trt. Mean	Ctrl. Mean	P (bootstrapping)
High Income	Fresh F&V Consumption	3.265	0.631	0.652	< 0.001
High Income	Fast Food Consumption	-6.772	0.367	0.342	< 0.001
High Income	Soda Consumption	-5.219	0.036	0.034	< 0.001
High Income	BMI	-0.335	28.077	27.983	< 0.001
High Income	% Overweight or Obese	0.643	0.647	0.651	0.006
High Education	Fresh F&V Consumption	9.180	0.602	0.657	< 0.001
High Education	Fast Food Consumption	-8.457	0.371	0.340	< 0.001
High Education	Soda Consumption	-10.600	0.038	0.034	< 0.001
High Education	BMI	-5.053	29.266	27.787	< 0.001
High Education	% Overweight or Obese	-13.100	0.734	0.638	< 0.001
Low Fast Food	Fresh F&V Consumption	5.311	0.612	0.645	< 0.001
Low Fast Food	Fast Food Consumption	-6.176	0.370	0.347	< 0.001
Low Fast Food	Soda Consumption	-15.725	0.039	0.033	< 0.001
Low Fast Food	BMI	-0.335	28.467	28.371	< 0.001
Low Fast Food	% Overweight or Obese	-1.474	0.679	0.669	< 0.001
High Grocery	Fresh F&V Consumption	3.437	0.606	0.627	< 0.001
High Grocery	Fast Food Consumption	-7.581	0.390	0.361	< 0.001
High Grocery	Soda Consumption	-8.567	0.039	0.036	< 0.001
High Grocery	BMI	-0.698	28.839	28.638	< 0.001
High Grocery	% Overweight or Obese	-2.437	0.699	0.682	< 0.001
Low Countertop Svc.	Fresh F&V Consumption	-0.177	0.612	0.610	0.289
Low Countertop Svc.	Fast Food Consumption	-0.190	0.384	0.384	0.337
Low Countertop Svc.	Soda Consumption	0.467	0.040	0.040	0.270
Low Countertop Svc.	BMI	0.102	28.710	28.740	0.082
Low Countertop Svc.	% Overweight or Obese	0.012	0.690	0.690	0.499
Low Electronics Stores	Fresh F&V Consumption	-0.681	0.609	0.605	0.011
Low Electronics Stores	Fast Food Consumption	-0.482	0.387	0.386	0.118
Low Electronics Stores	Soda Consumption	-0.406	0.040	0.040	0.278
Low Electronics Stores	BMI	0.006	28.906	28.908	0.467
Low Electronics Stores	% Overweight or Obese	0.018	0.701	0.701	0.467
Low Waterproofing Svc.	Fresh F&V Consumption	-1.607	0.614	0.604	< 0.001
Low Waterproofing Svc.	Fast Food Consumption	-1.319	0.389	0.384	< 0.001
Low Waterproofing Svc.	Soda Consumption	0.600	0.040	0.040	0.222
Low Waterproofing Svc.	BMI	0.046	28.786	28.799	0.274
Low Waterproofing Svc.	% Overweight or Obese	0.239	0.694	0.695	0.133

Table e4: Effect sizes of all race-specific top/bottom half matching experiments (Fig. 5).

Race	Treatment	Outcome	% Change	Trt. Mean	Ctrl. Mean	P-value (bootstrapping)
Black	High Income	Fresh F&V Consumption	-6.497	0.607	0.567	0.004
Black	High Income	Fast Food Consumption	5.464	0.403	0.425	0.015
Black	High Income	Soda Consumption	11.781	0.033	0.037	0.081
Black	High Income	BMI	3.695	29.834	30.936	< 0.001
Black	High Income	% Overweight or Obese	8.101	0.750	0.811	< 0.001
Black	High Education	Fresh F&V Consumption	11.243	0.558	0.620	< 0.001
Black	High Education	Fast Food Consumption	-7.613	0.424	0.391	0.002
Black	High Education	Soda Consumption	6.088	0.033	0.035	0.150
Black	High Education	BMI	-5.512	31.362	29.634	< 0.001
Black	High Education	% Overweight or Obese	-11.481	0.829	0.734	< 0.001
Black	Low Fast Food	Fresh F&V Consumption	7.024	0.543	0.582	< 0.001
Black	Low Fast Food	Fast Food Consumption	-12.043	0.473	0.416	< 0.001
Black	Low Fast Food	Soda Consumption	-14.421	0.040	0.034	< 0.001
Black	Low Fast Food	BMI	0.506	30.776	30.931	0.153
Black	Low Fast Food	% Overweight or Obese	3.062	0.775	0.799	0.001
Black	High Grocery	Fresh F&V Consumption	10.230	0.527	0.581	< 0.001
Black	High Grocery	Fast Food Consumption	-12.642	0.478	0.418	< 0.001
Black	High Grocery	Soda Consumption	-9.300	0.039	0.036	0.002
Black	High Grocery	BMI	-3.795	31.966	30.753	< 0.001
Black	High Grocery	% Overweight or Obese	-8.960	0.861	0.783	< 0.001
Hispanic	High Income	Fresh F&V Consumption	5.706	0.556	0.588	0.012
Hispanic	High Income	Fast Food Consumption	-3.314	0.393	0.380	0.140
Hispanic	High Income	Soda Consumption	0.663	0.036	0.036	0.429
Hispanic	High Income	BMI	0.289	28.997	29.080	0.280
Hispanic	High Income	% Overweight or Obese	-0.039	0.722	0.722	0.492
Hispanic	High Education	Fresh F&V Consumption	8.859	0.575	0.626	< 0.001
Hispanic	High Education	Fast Food Consumption	-11.902	0.385	0.339	< 0.001
Hispanic	High Education	Soda Consumption	-8.530	0.034	0.031	0.061
Hispanic	High Education	BMI	-5.949	29.738	27.969	< 0.001
Hispanic	High Education	% Overweight or Obese	-13.697	0.758	0.654	< 0.001
Hispanic	Low Fast Food	Fresh F&V Consumption	1.501	0.562	0.570	0.083
Hispanic	Low Fast Food	Fast Food Consumption	-5.920	0.423	0.398	< 0.001
Hispanic	Low Fast Food	Soda Consumption	-0.538	0.038	0.038	0.433
Hispanic	Low Fast Food	BMI	-0.172	29.750	29.699	0.291
Hispanic	Low Fast Food	% Overweight or Obese	-1.800	0.763	0.750	0.004
Hispanic	High Grocery	Fresh F&V Consumption	7.351	0.525	0.563	< 0.001
Hispanic	High Grocery	Fast Food Consumption	-7.212	0.443	0.411	< 0.001
Hispanic	High Grocery	Soda Consumption	-1.956	0.040	0.039	0.178
Hispanic	High Grocery	BMI	-1.518	30.358	29.898	< 0.001
Hispanic	High Grocery	% Overweight or Obese	-3.492	0.791	0.763	< 0.001
White	High Income	Fresh F&V Consumption	2.182	0.643	0.657	< 0.001
White	High Income	Fast Food Consumption	-5.063	0.359	0.341	< 0.001
White	High Income	Soda Consumption	-5.512	0.037	0.035	< 0.001
White	High Income	BMI	0.486	27.759	27.894	< 0.001
White	High Income	% Overweight or Obese	3.261	0.625	0.646	< 0.001
White	High Education	Fresh F&V Consumption	9.690	0.600	0.658	< 0.001
White	High Education	Fast Food Consumption	-5.878	0.363	0.342	< 0.001
White	High Education	Soda Consumption	-4.346	0.036	0.035	< 0.001
White	High Education	BMI	-4.100	28.955	27.768	< 0.001
White	High Education	% Overweight or Obese	-11.105	0.717	0.638	< 0.001
White	Low Fast Food	Fresh F&V Consumption	6.028	0.625	0.663	< 0.001
White	Low Fast Food	Fast Food Consumption	-6.611	0.359	0.335	< 0.001
White	Low Fast Food	Soda Consumption	-16.363	0.039	0.033	< 0.001
White	Low Fast Food	BMI	-0.644	28.147	27.966	< 0.001
White	Low Fast Food	% Overweight or Obese	-2.373	0.662	0.647	< 0.001
White	High Grocery	Fresh F&V Consumption	1.655	0.634	0.644	< 0.001
White	High Grocery	Fast Food Consumption	-5.004	0.368	0.349	< 0.001
White	High Grocery	Soda Consumption	-5.319	0.038	0.036	< 0.001
White	High Grocery	BMI	-0.104	28.239	28.209	0.082
White	High Grocery	% Overweight or Obese	-1.373	0.665	0.656	< 0.001

Table e5: Effect sizes of all top/bottom quartiles matching experiments (Fig. e1).

Treatment	Outcome	% Change	Trt. Mean	Ctrl. Mean	P-value (bootstrapping.)
High Income	Fresh F&V Consumption	7.383	0.636	0.683	< 0.001
High Income	Fast Food Consumption	-9.366	0.345	0.312	< 0.001
High Income	Soda Consumption	-7.064	0.032	0.029	< 0.001
High Income	BMI	0.689	27.175	27.362	< 0.001
High Income	% Overweight or Obese	2.904	0.596	0.613	< 0.001
High Education	Fresh F&V Consumption	13.977	0.582	0.664	< 0.001
High Education	Fast Food Consumption	-10.323	0.382	0.342	< 0.001
High Education	Soda Consumption	-12.293	0.039	0.034	< 0.001
High Education	BMI	-6.641	29.510	27.550	< 0.001
High Education	% Overweight or Obese	-15.904	0.739	0.621	< 0.001
High Grocery	Fresh F&V Consumption	5.664	0.614	0.649	< 0.001
High Grocery	Fast Food Consumption	-11.619	0.379	0.335	< 0.001
High Grocery	Soda Consumption	-14.903	0.038	0.032	< 0.001
High Grocery	BMI	-0.646	28.564	28.379	< 0.001
High Grocery	% Overweight or Obese	-2.613	0.683	0.665	< 0.001
Low Fast Food	Fresh F&V Consumption	8.893	0.613	0.668	< 0.001
Low Fast Food	Fast Food Consumption	-11.542	0.371	0.328	< 0.001
Low Fast Food	Soda Consumption	-30.282	0.042	0.029	< 0.001
Low Fast Food	BMI	-0.153	28.161	28.118	0.098
Low Fast Food	% Overweight or Obese	-0.384	0.656	0.654	0.140

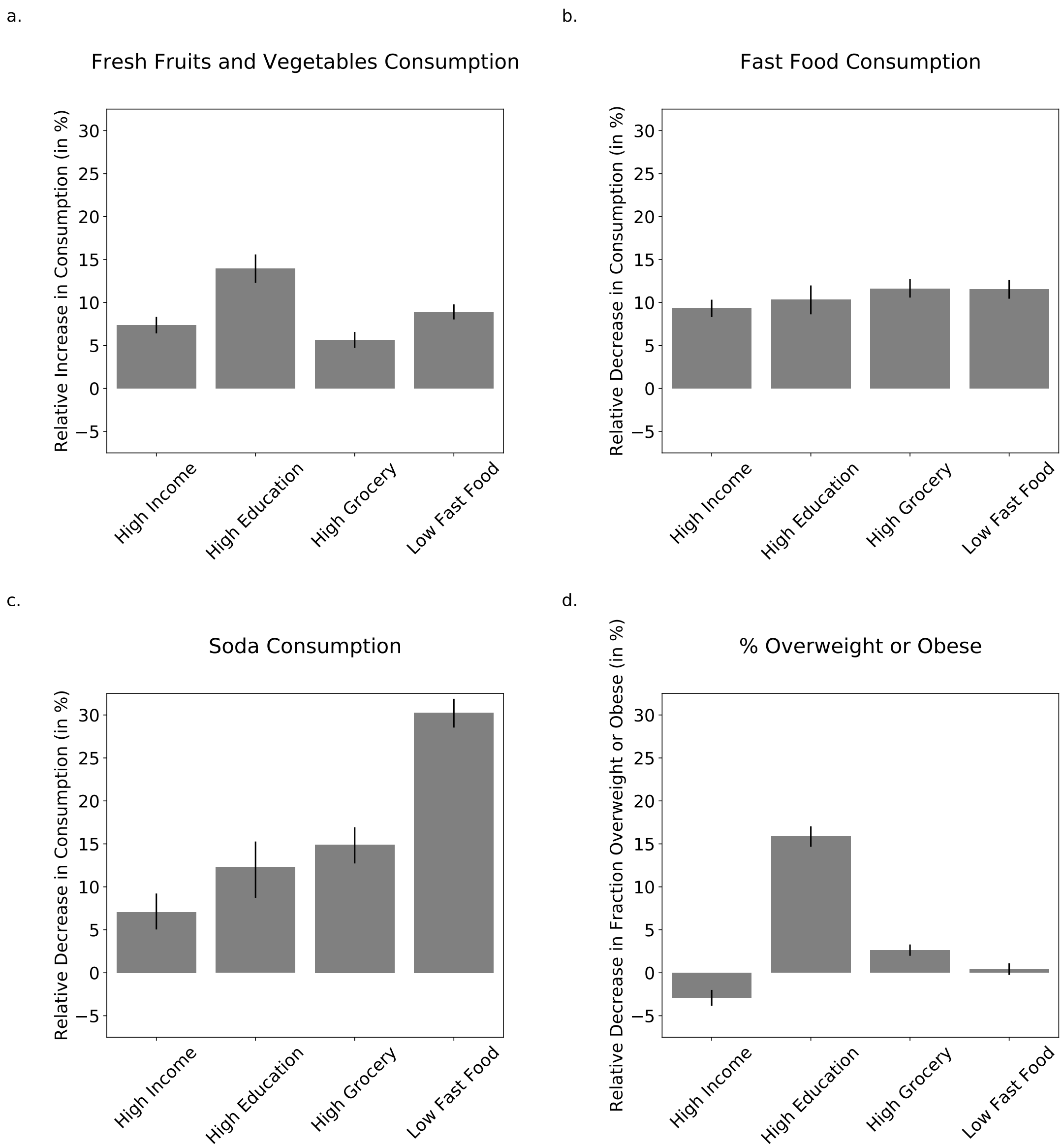

Figure e1: Matching experiments using quartile instead of median split. Note the consistent but increased effect sizes compared to Figure 4.

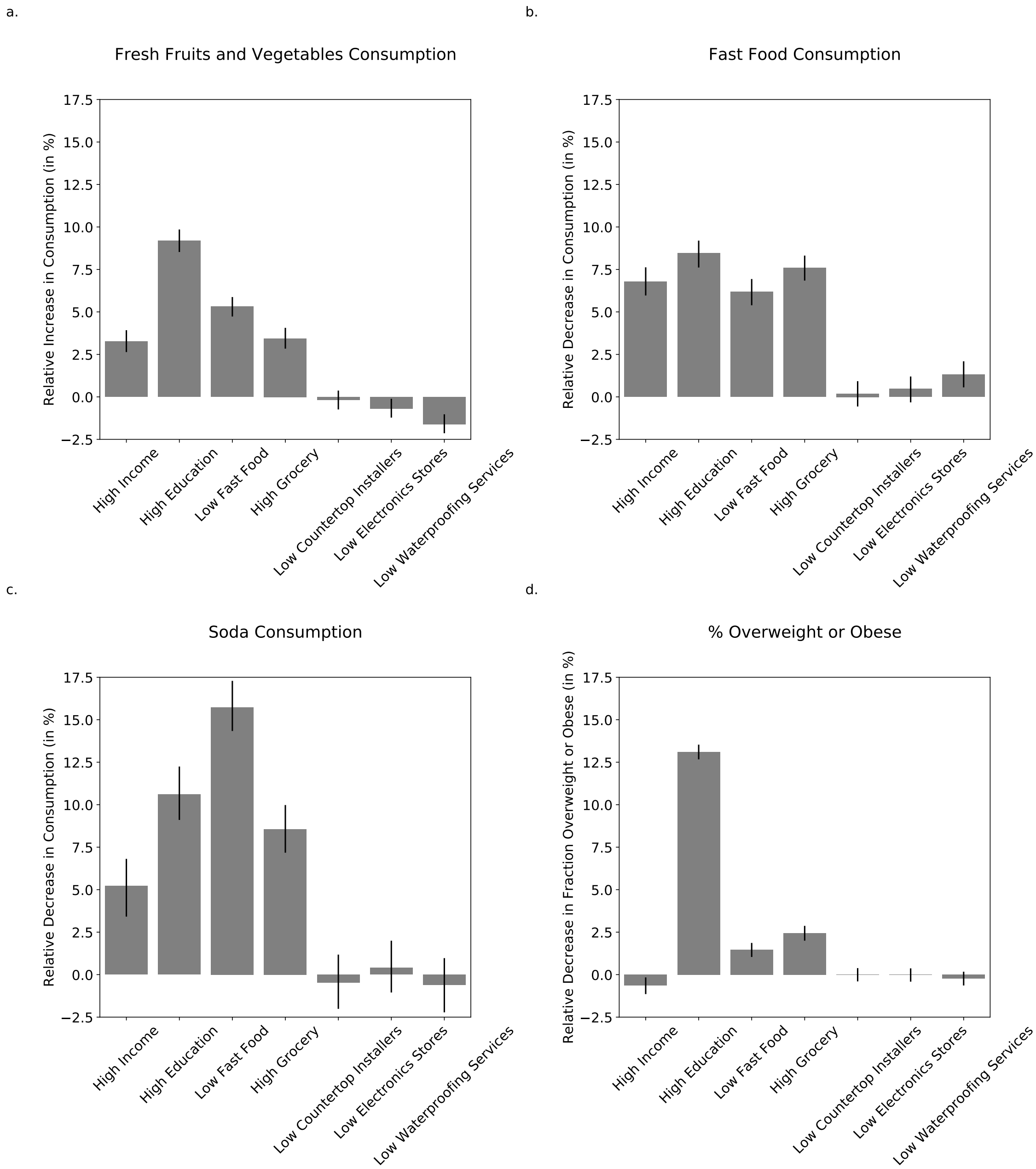

Figure e2: **Demonstration of discriminant validity of statistical approach.** We measured the effect of null-treatments that should not have any impact on food consumption. We chose examples of null-treatments by selecting variables that had little correlation with study independent variables (income, education, grocery access, fast food access) and were plausibly unrelated to food consumption. This selection process lead to use of the fraction of countertop installers, electronics stores, and waterproofing services nearby as measured through Yelp. Applying our analysis pipeline to these null-treatments, we found that all of these effect estimates were close to zero. This demonstrated that our statistical analysis approach did not produce measurements that it was not supposed to measure; that is, discriminant validity.

Table e6: USA fast food restaurants table used to classify participant food entries as fast food[102]. A list of popular pizza chains from the USA was appended to the list[103].

A&W Restaurants	Cinnabon	Red Burrito	Rogers Restaurants	Chuck E. Cheese's	Murphy's Pat's Pizza
Arby's	Claim Jumper	The Habit	Runza Saladworks	CiCi's Pizza Cottage	Patxi's Chicago Peter
Arctic Circle	Coco's	Halal Guys	Sbarro Schlotzsky's	Inn Dion's Discovery	Piper Pie Five
Arthurs	Cold Stone Creamery	Hardee's	Seattle's Best Shake	Zone Domino's	Pietro's Pizza Pizza
Atlanta Bread	Cookout	Huddle House	Shack Skyline Chili	Donatos	Corner Pizza Factory
Au Bon Pain	Copeland's	In-N-Out Burger	Sneaky Pete's Sonic	DoubleDave's East of	Pizza Fusion Pizza
Auntie Anne's	Old Country	Jack in the Box	Spangles Steak	Chicago Eatza	Hut Pizza King Pizza
Baja Fresh	Culver's	Jack's Family	Escape Steak 'n	Extreme Pizza	Inn Pizza My Heart
Bakers Square	Dairy Queen	Restaura	Shake Stir Crazy Sub	Fazoli's Fellini's	Pizza Patrón Pizza
Blimpies	el Tacos	ts Jersey Mike's Subs	Station II Subway	Fox's Frank Pepe	Ranch Pizza
Bojangles	DiBella's	Jimmy John's Jim's	Swensen's Swensons	Gatti's Gino's	Schmizza The Pizza
Boston Market	Dixie Chili and Deli	Restaurants Johnny	Taco Bell Taco	Giordano's	Studio Pizzeria Venti
Braum's	Don Pablo's	Rockets KFC	Bueno Taco Cabana	Godfather's	Regina Pizzeria
Burger Chef	Druther's	Kewpee Krispy	Taco John's Taco	Grimaldi's Grotto	Rocky Rococo
Burger King	Dunkin' Donuts	Kreme L&L	Mayo Taco Tico Taco	Pizza Happy Joe's	Rosati's Round Table
Burger Street	Eat'n Park	Hawaiian Barbecue	Time Twin Peaks	Happy's Hideaway	Pizza Russo's New
Burgerville	Eegee's	Lee Roy Selmon's	Umami Burger	Home Run Inn	York Pizze
Captain D's	El Chico	Lee's Famous Lion's	Wendy's Wetzel's	Hungry Howie's	ia Sal's Pizza
Carino's Italian Grill	El Pollo Loco	Choice Long John	Pretzels Whataburger	Hunt Brothers Imo's	Sammy's
Carl's Jr.	El Taco Tote	Silver's Luby's	White Castle	Pizza Jerry's Jet's	Sarpino's
Carrows	Elephant Bar	McDonald's Milo's	Wienerschnitzel	John's LaRosa's	Sbarro
Charley's Grilled	Elevation Burger	Moe's Mooyah Mr.	Wimpy Zaxby's	Ledo Little Caesars	Shakey's
Subs	Famous Dave's	Hero Mrs. Fields	Zero's Subs Zippy's	Lou Malnati's	Showbiz
Checkers	Farmer Boys	Mrs. Winner's	America's Incredible	Marco's Marion's	Sir Pizza
Cheeburger	Fatburger	Chicken	Arni's Aurelio's	Mark's Mazzio's	Snappy Tomato
Cheeburger	Firehouse Subs	Biscuits Naugles	Azzip Bearno's	Mellow Mushroom	Straw Hat
Chevys	Five Guys	Panera Bread Panda	Bertucci's Big	MOD Pizza	Toppers
Chicken Express	Freddy's	Express Penn Station	Mama's & Papa's	Monical's Mountain	Uncle Maddio's
Chick-fil-A	Freddies	Pita Pit Popeyes Port	Blackjack Blaze	Mike's Mr. Jim's	Unos
Chronic Tacos	Golden Chick	of Subs Potbell	Buddy's Bullwinkle's	Noble Roman's Old	Upper Crust Pizzeria
Chuck-A-Rama	Good Times	Quizno's Raising	California Pizza	Chicago Pacpizza	Valentino's
Church's	Great Steak	Cane's Rax Roast	Casey's General	Pagliacci Papa Gino's	Vocelli Pizza
Texas Chicken	Green Burrito	Beef Robeks Roy	Stores Cassano's	Papa John's Papa	Your Pie

Table e7: USA soda list used to classify participant food entries as sugary sodas. The list was constructed using a list of America's best-selling brands of Soda[104], in addition to the generic terms such "Root Beer" and "Coca".

Pepsi	Dr Pepper	Root Beer
Coca Cola	Sprite	Coke
Mountain Dew	Fanta	Coca

Table e8: Summary of High Income (MedianFamilyIncome > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4911	4911
Matched	1358	4911
Unmatched	3553	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.75	0.75	0.26	0.00	0.01
Education (% Without College Degree)	0.56	0.57	0.15	-0.01	-0.06
Grocery Distance (USDA lapophalfshare)	0.74	0.74	0.21	0.00	0.00
Yelp Fast Food %	0.06	0.06	0.06	0.00	-0.01
Median Family Income	97244.90	58991.91	8680.39	38253.00	4.41

Table e9: Summary of High Grocery (Grocery Access > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4911	4911
Matched	2219	4911
Unmatched	2692	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.59	0.58	0.19	0.00	0.01
Median Family Income	76817.91	77021.36	30133.94	-203.45	-0.01
Education (% Without College Degree)	0.64	0.64	0.18	0.00	-0.01
Yelp Fast Food %	0.06	0.06	0.06	0.00	0.00
Grocery Distance (USDA lapophalfshare)	0.57	0.88	0.06	-0.31	-5.62

Table e10: Summary of High Education (% College Degrees > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4911	4911
Matched	1481	4911
Unmatched	3430	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.73	0.73	0.27	0.00	0.01
Median Family Income	93203.36	88493.86	20530.36	4709.51	0.23
Grocery Distance (USDA lapophalfshare)	0.71	0.72	0.22	-0.01	-0.05
Yelp Fast Food %	0.06	0.06	0.06	0.00	0.02
Education (% Without College Degree)	0.53	0.75	0.05	-0.22	-4.79

Table e11: Summary of Low Fast Food (% Yelp Fast Food < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4911	4911
Matched	1910	4911
Unmatched	3001	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.65	0.65	0.24	0.00	0.01
Median Family Income	86942.10	86597.30	31088.00	344.80	0.01
Education (% Without College Degree)	0.60	0.61	0.18	0.00	-0.02
Grocery Distance (USDA lapophalfshare)	0.66	0.67	0.23	-0.01	-0.05
Yelp Fast Food %	0.03	0.11	0.06	-0.08	-1.35

Table e12: Summary of High Income (MedianFamilyIncome > 75th Percentile) matching experiment

(a) Sample sizes

	Control	Treated
All	2319	2321
Matched	215	2321
Unmatched	2104	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.91	0.91	0.18	0.00	0.01
Education (% Without College Degree)	0.47	0.48	0.13	-0.02	-0.11
Grocery Distance (USDA lapophalfshare)	0.71	0.70	0.22	0.01	0.04
Yelp Fast Food %	0.04	0.04	0.04	0.00	-0.06
Median Family Income	115443.80	48357.34	6021.64	67086.46	11.14

Table e13: Summary of High Grocery (Grocery Access > 75th Percentile) matching experiment

(a) Sample sizes

	Control	Treated
All	2321	2319
Matched	743	2319
Unmatched	1578	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.69	0.68	0.23	0.01	0.03
Median Family Income	77354.53	77809.00	31756.28	-454.47	-0.01
Education (% Without College Degree)	0.62	0.62	0.19	0.00	-0.01
Yelp Fast Food %	0.04	0.04	0.04	0.00	-0.01
Grocery Distance (USDA lapophalfshare)	0.41	0.94	0.04	-0.52	-13.78

Table e14: Summary of High Education (% College Degrees > 75th Percentile) matching experiment. **Note:** Treatment samples dropped due to 0.3 STD caliper used to ensure 0.25 SMD balancing constraint.

(a) Sample sizes

	Control	Treated
All	2321	2319
Matched	252	834
Unmatched	2069	1485
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.73	0.71	0.30	0.02	0.06
Median Family Income	81374.28	79453.49	12961.32	1920.79	0.15
Grocery Distance (USDA lapophalfshare)	0.70	0.70	0.22	0.00	-0.01
Yelp Fast Food %	0.06	0.06	0.06	0.00	0.02
Education (% Without College Degree)	0.48	0.83	0.04	-0.34	-9.25

Table e15: Summary of Low Fast Food (% Yelp Fast Food < 25th Percentile) matching experiment. **Note:** Treatment samples dropped due to 1.5 STD caliper used to ensure 0.25 SMD balancing constraint.

(a) Sample sizes

	Control	Treated
All	2322	2319
Matched	504	2166
Unmatched	1818	153
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.78	0.77	0.27	0.01	0.04
Median Family Income	88631.94	83632.13	20794.81	4999.81	0.24
Education (% Without College Degree)	0.59	0.59	0.15	0.00	-0.02
Grocery Distance (USDA lapophalfshare)	0.63	0.66	0.23	-0.03	-0.14
Yelp Fast Food %	0.02	0.18	0.05	-0.16	-3.23

Table e16: Summary of Low Countertop Installation Services (% Yelp Countertop Installers < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4913	4909
Matched	2829	4909
Unmatched	2084	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.50	0.50	0.05	0.00	0.01
Median Family Income	77255.32	77443.29	28361.82	-187.96	-0.01
Education (% Without College Degree)	0.65	0.66	0.18	0.00	0.00
Grocery Distance (USDA lapophalfshare)	0.72	0.72	0.24	0.00	0.00
Yelp Fast Food %	0.09	0.09	0.08	0.00	0.01
Yelp Countertop Installers %	0.00	0.00	0.00	0.00	-1.61

Table e17: Summary of Low Electronics Stores (% Yelp Electronics Stores < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4912	4910
Matched	2741	4910
Unmatched	2171	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.52	0.52	0.09	0.00	0.00
Median Family Income	74567.18	74306.84	25820.76	260.34	0.01
Education (% Without College Degree)	0.68	0.68	0.16	0.00	0.00
Grocery Distance (USDA lapophalfshare)	0.71	0.71	0.22	0.00	-0.01
Yelp Fast Food %	0.08	0.08	0.07	0.00	0.00
Yelp Electronics Stores %	0.00	0.00	0.00	0.00	-1.59

Table e18: Summary of Low Waterproofing Services (% Yelp Waterproofing Services < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4913	4909
Matched	2876	4909
Unmatched	2037	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.50	0.50	0.02	0.00	0.00
Median Family Income	76371.54	76438.89	28071.29	-67.35	0.00
Education (% Without College Degree)	0.66	0.66	0.17	0.00	0.00
Grocery Distance (USDA lapophalfshare)	0.74	0.74	0.21	0.00	-0.01
Yelp Fast Food %	0.09	0.09	0.08	0.00	0.01
Yelp Waterproofing Services %	0.00	0.00	0.00	0.00	-1.48

Table e19: Summary of Black-majority Zip Code High Income (MedianFamilyIncome > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	317	42
Matched	30	42
Unmatched	287	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.36	0.35	0.25	0.02	0.07
Education (% Without College Degree)	0.65	0.65	0.09	-0.01	-0.08
Grocery Distance (USDA lapophalfshare)	0.68	0.67	0.22	0.01	0.02
Yelp Fast Food %	0.04	0.03	0.03	0.00	0.09
Median Family Income	88944.14	59779.41	8039.94	29164.73	3.63

Table e20: Summary of Hispanic-majority Zip Code High Income (MedianFamilyIncome > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	482	67
Matched	51	67
Unmatched	431	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.35	0.31	0.25	0.04	0.16
Education (% Without College Degree)	0.70	0.72	0.12	-0.02	-0.16
Grocery Distance (USDA lapophalfshare)	0.62	0.63	0.20	-0.01	-0.05
Yelp Fast Food %	0.07	0.07	0.05	0.00	0.00
Median Family Income	82812.73	56050.33	8534.43	26762.40	3.14

Table e21: Summary of White-majority Zip Code High Income (MedianFamilyIncome > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	3421	4277
Matched	1023	4277
Unmatched	2398	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.78	0.78	0.25	0.00	0.01
Education (% Without College Degree)	0.55	0.56	0.15	-0.01	-0.07
Grocery Distance (USDA lapophalfshare)	0.76	0.76	0.19	0.00	-0.01
Yelp Fast Food %	0.06	0.06	0.06	0.00	-0.01
Median Family Income	98014.79	59878.82	8719.97	38135.97	4.37

Table e22: Summary of Black-majority Zip Code High Grocery (Grocery Access > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	100	259
Matched	65	259
Unmatched	35	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.78	0.77	0.15	0.01	0.08
Median Family Income	51410.82	51925.49	14357.87	-514.66	-0.04
Education (% Without College Degree)	0.76	0.77	0.08	0.00	-0.03
Yelp Fast Food %	0.05	0.05	0.05	0.00	-0.08
Grocery Distance (USDA lapophalfshare)	0.53	0.86	0.04	-0.33	-9.43

Table e23: Summary of Hispanic-majority Zip Code High Grocery (Grocery Access > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	78	471
Matched	66	471
Unmatched	12	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.87	0.87	0.09	0.00	0.01
Median Family Income	52218.86	53068.49	13918.72	-849.63	-0.06
Education (% Without College Degree)	0.82	0.83	0.08	-0.01	-0.12
Yelp Fast Food %	0.06	0.06	0.05	0.00	0.01
Grocery Distance (USDA lapophalfshare)	0.47	0.88	0.05	-0.41	-7.86

Table e24: Summary of White-majority Zip Code High Grocery (Grocery Access > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4494	3204
Matched	1741	3204
Unmatched	2753	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.51	0.50	0.19	0.00	0.02
Median Family Income	84030.48	84275.98	31002.90	-245.50	-0.01
Education (% Without College Degree)	0.59	0.59	0.18	0.00	-0.01
Yelp Fast Food %	0.07	0.07	0.07	0.00	0.00
Grocery Distance (USDA lapophalfshare)	0.62	0.89	0.05	-0.27	-4.92

Table e25: Summary of Black-majority Zip Code High Education (% College Degrees > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	285	74
Matched	48	74
Unmatched	237	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.46	0.45	0.29	0.01	0.02
Median Family Income	70932.69	71048.45	22123.91	-115.77	-0.01
Grocery Distance (USDA lapophalfshare)	0.59	0.60	0.28	-0.01	-0.04
Yelp Fast Food %	0.04	0.04	0.04	0.00	0.00
Education (% Without College Degree)	0.62	0.77	0.04	-0.15	-3.39

Table e26: Summary of Hispanic-majority Zip Code High Education (% College Degrees > Median) matching experiment

(a) Sample sizes

	Control	Treated
All	488	61
Matched	43	61
Unmatched	445	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.33	0.28	0.22	0.05	0.21
Median Family Income	71569.13	67609.51	16348.80	3959.62	0.24
Grocery Distance (USDA lapophalfshare)	0.51	0.52	0.27	-0.01	-0.04
Yelp Fast Food %	0.05	0.05	0.04	0.00	-0.06
Education (% Without College Degree)	0.61	0.80	0.05	-0.19	-3.79

Table e27: Summary of White-majority Zip Code High Education (% College Degrees > Median) matching experiment. **Note:** Treatment samples dropped due to 2.1 STD caliper used to ensure 0.25 SMD balancing constraint.

(a) Sample sizes

	Control	Treated
All	3491	4207
Matched	1114	4102
Unmatched	2377	105
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.75	0.75	0.26	0.00	0.01
Median Family Income	92654.95	88357.13	17926.24	4297.82	0.24
Grocery Distance (USDA lapophalfshare)	0.74	0.76	0.17	-0.01	-0.07
Yelp Fast Food %	0.07	0.07	0.06	0.00	0.00
Education (% Without College Degree)	0.53	0.75	0.05	-0.22	-4.59

Table e28: Summary of Black-majority Zip Code Low Fast Food (% Yelp Fast Food < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	100	259
Matched	70	259
Unmatched	30	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.77	0.76	0.16	0.00	0.03
Median Family Income	54812.73	54366.39	18020.16	446.34	0.02
Education (% Without College Degree)	0.75	0.74	0.10	0.01	0.09
Grocery Distance (USDA lapophalfshare)	0.57	0.60	0.22	-0.02	-0.11
Yelp Fast Food %	0.03	0.11	0.07	-0.08	-1.21

Table e29: Summary of Hispanic-majority Zip Code Low Fast Food (% Yelp Fast Food < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	252	297
Matched	135	297
Unmatched	117	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.61	0.61	0.19	0.00	0.02
Median Family Income	52130.57	51853.43	13411.40	277.14	0.02
Education (% Without College Degree)	0.81	0.82	0.09	-0.01	-0.11
Grocery Distance (USDA lapophalfshare)	0.45	0.45	0.26	0.00	-0.01
Yelp Fast Food %	0.03	0.10	0.06	-0.07	-1.20

Table e30: Summary of White-majority Zip Code Low Fast Food (% Yelp Fast Food < Median) matching experiment

(a) Sample sizes

	Control	Treated
All	4188	3510
Matched	1362	3510
Unmatched	2826	0
Discarded	0	0

(b) Summary of balance for matched data

	Means Treated	Means Control	SD Control	Mean Diff	Std. Mean Difference
distance	0.65	0.65	0.26	0.00	0.01
Median Family Income	95359.35	94854.08	29365.91	505.27	0.02
Education (% Without College Degree)	0.56	0.57	0.17	0.00	-0.01
Grocery Distance (USDA lapophalfshare)	0.71	0.72	0.20	-0.01	-0.05
Yelp Fast Food %	0.03	0.11	0.06	-0.08	-1.37