Changes in Household Diet: Determinants and Predictability *

Stefan Hut
Brown University

Emily Oster
Brown University and NBER

January 15, 2018

Abstract

American diets are, on average, unhealthy relative to nutritional guidelines. We combine household food purchase data with health information and the timing of diet-related news releases to estimate what types of information or education improves diet and for whom. Our primary finding is that the average household does not improve diet quality in response to serious disease diagnosis (diabetes, hypertension, obesity), changes in government diet recommendations, or major research findings. Households with more economic incentives to be healthy - higher income, higher education, younger - also do not respond significantly. This suggests that dietary choices are difficult to alter. We identify households in the panel who do show large improvements in their diet quality; this is 5 to 6% of the sample. We use a machine learning approach to predicting these households, and identify that the concentration of baseline diet - the extent to which it loads on a small number of categories - is a strong predictor of diet quality improvement. We comment on policy.

1 Introduction

What causes people to improve the quality of their diet? Are some people more likely to improve their diet than others? Can these individuals be predicted?

These questions have policy implications. At least two-thirds of American adults are estimated to be overweight, and a third are obese.[1] Obesity, and related conditions, are expensive for the health care system and have both morbidity and mortality consequences for individuals (Feldstein et al., 2008). It seems clear that poor diet plays a major role in driving differences in obesity rates, both in the cross section and over time (Cutler et al., 2003, Bleich et al., 2008, Swinburn et al., 2009). Improving the overall diet of Americans would, therefore, have both health and budgetary consequences. Perhaps in recognition of this, the US government spends significant funds on promoting healthy diets. Nutrition education is included in all the major government dietary programs, including school lunch plans, the Women, Infants and Children (WIC)

*We are grateful to Geoffrey Kocks for exceptional research assistance, as well as to Sofia La Porta, Julian De Georgia and Cathy Yue Bai. The conclusions drawn from the Nielsen data are those of the researchers and do not reflect the views of Nielsen. Nielsen is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein. Results are calculated based on data from The Nielsen Company (US), LLC and marketing databases provided by the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business.

[1] https://www.cdc.gov/obesity/data/adult.html

program and the Supplemental Nutrition Assistance Program (SNAP). In 2017, 414 million dollars were spent on nutrition education for participants in SNAP alone.[2] There is literature within economics which suggests that education programs may be key - rather than, say, grocery access - to combating obesity (e.g. Alcott, Diamond and Dube, 2017).

This education-focused approach presumes the healthfulness of diet is malleable. But there is at the same time other evidence suggesting dietary choices in general are persistent in the face of changes in circumstance or food options (Handbury et al, 2015; Bronnenberg, Dube and Gentzkow, 2012; Alcott et al, 2017; Atkin, 2016). Atkin (2016) shows, for example, that people are willing to give up valuable nutrients to maintain their preferred diet. To the extent that this force dominates the health incentives to change behavior, information and other nudges to encourage behavior change may be fruitless. However, observing that diet is stable in some situations does not rule out that other changes in circumstance or information will alter diet. Further, there may be some individuals whose diet is more malleable than others, and this may not be captured in averages.

This paper takes a data-driven approach to the questions posed at the start. We use household scanner data - from the Nielsen HomeScan Panel - to observe a measure of the quality of diet in a large panel of households over time. This allows us to analyze changes in diet in a population which is not involved in a targeted dietary intervention. The paper proceeds in three parts.

First, we use an event study design to estimate changes in the quality of diet in response to a variety of events which could prompt dietary changes: individual diagnosis of chronic disease, information from government diet education policies, and research findings. In this analysis we also estimate heterogeneity in response across demographics. Broadly, we find little or no response on average to any of these events, and no demographic groups with meaningful changes.

Second, we use the data to search more generally for any households in the panel that show substantial, sustained improvements in diet quality. We find that 5 to 6 percent of households show large dietary improvements at some point in time.

Third, we ask to what extent these large-change households can be predicted *ex ante*. We show that these households are only weakly predicted with demographics and basic baseline diet features. We find, however, that using a machine learning approach can substantially improve prediction. Although the prediction process of the algorithm is complex, we show that we can use the output to identify some central predictors of behavior change. The features identified provide some insight into what may facilitate dietary improvements.

We begin in Section 2 by describing the Nielsen HomeScan database, which is key to the analysis in this paper. This is a household-based dataset in which participant households scan grocery (and other) purchases with a home hand scanner. We use these data to construct a monthly panel measuring the healthfulness of household grocery purchases. Nielsen data contains detailed product information in the form of UPC codes, but does not contain nutrition information directly. We merge the purchase data with nutrition information

[2] http://www.obpa.usda.gov/budsum/fy17budsum.pdf

gathered from a number of sources, notably from the USDA.

We focus on three measures of diet quality: (1) the share of purchases in soda and sugar-based categories; (2) the share of purchases in vegetables, fruits and whole grains and (3) the nutrient ratio, which is a composite measure based on the nutrients in all purchases. The panel contains about 160,000 households, observed for an average of 48 months.

There are some limitations to these data. Nielsen does not provide a full picture of realized diet (most notably it describes only purchases, not consumption, and excludes food away from home) and is observed at the household rather than the individual level. In addition, there is no exogenous treatment assignment for us to exploit. However, the structure of the data also provides some substantial advantages. Many households are in the panel for a long period, allowing a detailed look at diet over a long time. Further, the panel is large; when we get to the second step, where we seek to identify households with sustained dietary improvements, it will be advantageous to have a large sample. Finally, since the panel is not focused on dietary health, and does not recruit panelists based on either healthy or unhealthy diets, we have a window into changes in diet which occur in a general population outside of particular interventions.

We link the purchase data from HomeScan with an additional Nielsen survey called the Ailment Panel to identify diagnoses of metabolic disease among household members - this includes diabetes, hypertension and obesity. In addition, we add to the data information about the timing of a key government educational intervention (the move to a food plate instead of a pyramid) and about the release of results from several important diet studies.

In Section 3, we present our baseline analysis of household response to the events considered. We use a household fixed effect design to look for changes within a household over time around these events. We find very limited impacts of any of these events on diet quality. Diabetes diagnosis appears to result in some dietary improvements over the following two years, although these are small in magnitude. We see a small increase in purchases of nuts and olive oil after large study releases suggesting a Mediterranean diet supplemented with these foods reduced mortality. But these changes are tiny, roughly 0.01 of a standard deviation, and they persist for only a short time.

An analysis of heterogeneity in these patterns across demographics – age, education, income – also yields little. To the extent there is any heterogeneity, it seems to go the opposite direction of what we would expect; for example, less educated diabetics respond marginally more. The primary source of heterogeneity is in baseline diet: households with a worse baseline diet show more improvement. This is at least partially mechanical and may not be very helpful for targeting, other than to say that households with worse diets have more room to improve.

One interpretation of these initial findings is that diet is simply not malleable and there is no one who substantially improves the quality of their diet. It is also possible, however, that there are households or individuals who successfully improve their diet, but we have not identified either the events or the individual characteristics which correlate with this. These two explanations have differing policy implications.

To evaluate the plausibility of the second explanation, we turn in Section 4 to looking in the data to see if we can identify *any* households with sizable dietary improvements. We focus on a balanced panel of households which we consistently observe over a long period. We identify a set of households who show a sustained dietary improvement, as measured either by a reduction in the share of expenditure on unhealthy foods or by an improvement in the nutrient ratio measure. To limit concerns about regression to the mean, we require a long period of improved diet after a long baseline period.

Depending on the outcome, we identify 5 to 6% of the total balanced panel as significant changers. Excluding the period on which selection occurs, these households have sizable changes in diet: 25% to 30% of the mean and larger than the cross-sectional differences in diet quality across college versus non-college educated households.

Although this analysis suggests households with dietary improvements do exist, we find they are not well predicted by baseline demographics or summary features of their diet. This is consistent with our basic findings above.

In Section 5 we therefore turn to using a machine learning approach - in particular, a random forest algorithm - to predict these large-change households. Using a rich set of diet and household characteristics, we find that we are able to predict changes in diet with greater accuracy.

An issue with using machine learning in social science is that prediction is often not the goal (Kleinberg et al, 2017). To the extent we want to use this approach to learn what prompts behavior change, we need to be able to summarize the random forest output intuitively.

It is common to summarize the output of a random forest based on an importance plot which indicates which variables are most important in prediction. This summary, however, is not especially illuminating as it does not provide either a sense of the direction in which the features matter, or a sense of the importance of interactions between features.

We therefore adopt a visualization approach from Hastie et al (2009) and Jones and Linder (2015) to better illustrate the role of key features and the interactions between them. This approach allows us to show graphically the relationships between single variables and the outcome, and to illustrate two-way interactions. A key aspect of these visualizations is that the plots are partial dependence plots, not marginal plots, so they provide a descriptive sense of the relationship taking into account the other variables which move together with each feature.

The visualization generates two insights. First, we consistently observe that dietary concentration - as measured by the standard deviation of spending shares across food groups - is a strong predictor of change. The relationship is roughly linear: dietary improvement is more likely for households with a more concentrated diet *ex ante.* Second, diet concentration interacts with baseline diet: concentration is more predictive of change for households with a less healthy baseline diet.

Motivated by these findings, we return to our initial estimate of response to events - focusing on the disease diagnosis cases. We extract the key features and interactions identified in the random forest and

estimate heterogeneity in response along these features. We find there is much more consistent heterogeneity in response across these learning-identified features than across the demographics we consider initially.

From a policy standpoint, we argue that the results in this paper make two contributions. First, we find that on average the quality of diet is stagnant *even* in response to fairly substantial events which should prompt dietary improvement. This evidence is at least suggestive that providing information or encouragement to improve diet may not be sufficient.

Second, the results on diet concentration – as revealed by the learning analysis – suggest a new avenue to explore in thinking about the process behind dietary improvements, and in thinking about which households may be most responsive to interventions. One interpretation - although beyond what we can show in the data - is that a more concentrated diet lends itself to development of diet "rules" which can make behavior change easier.

This paper contributes to a large literature in public health and a smaller literature within economics about limitations to generating changes in diet (Handbury et al, 2015; Alcott et al, 2017; Delamater, 2006; Broadbent, Donkin and Stroh, 2011; Ponzo et al, 2017; Raj et al, 2017).We differ to some extent in our lack of focus on a particular intervention or treatment. We also contribute to a literature on minimal behavior change in response to educational interventions (e.g. Diabetes Prevention Group, 2009; Pi-Sunyer, 2014). Finally, we add to a literature within economics on how people respond to news about their health (Carrera, Heran & Prina, 2016; Oster, *forthcoming*).

We are also among a small (but growing) number of papers within economics which use machine learning techniques (e.g. Kleinberg et al, 2017; Oster, forthcoming; Gilchrist and Sands, 2016). The visualization approach we adopt here may be useful to others working in this space since it allows for a more intuitive description of results.

The rest of the paper proceeds as follows. Section 2 describes the data and Section 3 shows our baseline results. Section 4 describes the construction of the sample with large behavior changes, and Section 5 develops our machine learning approach. Section 6 concludes.

2 Data

This section describes the data used in the paper. The starting point of our empirical approach will be the ability to observe measures of diet over time. The Nielsen Homescan panel is described in Subsection 2.1. Information on the independent variables we consider appear in Subsection 2.2.

2.1 Diet Data

The primary data used in this paper is the Nielsen HomeScan panel. These data track consumer purchases using at-home scanner technology. Households that are part of the panel are asked to scan their purchases after all shopping trips; this includes grocery and pharmacy purchases, large retailer and super-center pur-

chases, as well as purchases made online and at smaller retailers. The Nielsen data records the UPC of items purchased and panelists provide information on the quantities, as well as information on the store. Prices are recorded by the panelists or drawn from Nielsen store-level data, where available. Einav, Leibtag and Nevo (2010) validate the reliability of the HomeScan panel. We use Nielsen data available through the Kilts Center at the University of Chicago Booth School of Business. These data cover purchases from 2004 through 2015.

Panel A of Table 1 shows some basic demographic features for the sample. The Nielsen data is intended to be a representative sample of the US on demographic features. In Appendix Table A1 we show the demographics for the sample used here relative to the US overall. The Nielsen sample is very similar on most demographics, although differs in the share white.

Our focus is on the healthfulness of diet. To analyze this, we need to convert the purchase behavior to metrics of diet quality. We rely in part on external data on the nutrition content of items at the UPC level. These data include the USDA Branded Food Products Database and the USDA National Nutrient Database for Standard Reference. We supplement this with nutrition facts for food items on Shopwell.com and Walmart.com and product nutrition facts from Labelinsight.com. Altogether, the data provides a direct UPC match for approximately 40% of UPCs and 75% of sales in the data. For the remaining items, we undertake an imputation procedure similar to that outlined in Dubois, Griffith and Nevo (2014). This procedure has two rounds. First, we impute the nutrition values using the average within product module, size type, brand, flavor, and formula (as defined by Nielsen). Many of the items that remain unmatched are store brand items. For these items we take the average within product module, size type, variety, type, formula, and style (i.e., drop the brand requirement). After imputation, approximately 7.6% of purchases in the data remain unmatched. We exclude these products from our calculation of the nutrient ratio.

Combining thesedata, we create several outcome measures designed to capture the healthfulness of diet. First, we construct two measures of diet quality based on expenditure share of groups defined in the USDA "Thirfty Food Plan". The TFP is one of four USDA-designed food plans specifying foods and amounts of foods that provide adequate nutrition. This approach to measuring diet quality has been used in related literature (Handbury, 2015; Volpe et al, 2013; Oster, 2017). First, we compute the share of total food spending on obvious unhealthful food categories, which contain soft drinks, soda, fruit drinks and ades, sugar, sweets and candy. Higher values for this indicate a less healthy diet. Second, we similarly calculate total expenditure share for obvious healthful food categories, containing vegetables, fruits and whole grains. Higher values for this indicate a healthier diet.

In addition to this, we use a summary measure of nutrition quality – the nutrient ratio. This index captures the extent to which a household's grocery purchases deviate from the nutrient composition recommended in the federal Dietary Guidelines for Americans (DGA). For each household h in month t, the nutrient ratio is defined as:

$$Nutrient\,Ratio_{ht} = \frac{\sum_{j \in J_{Healthful}} \left(\frac{pc_{jht}}{pc_j^{DGA}} \right)}{\sum_{j \in J_{Unhealthful}} \left(\frac{pc_{jht}}{pc_j^{DGA}} \right)}$$

where j indexes nutrients, pc_{jht} denotes the amount of nutrient j per calorie in household h's grocery purchases in month t, and pc_j^{DGA} is the amount of nutrient j in the DGA recommended diet per calorie consumed[3].

Nutrients in the unhealthful category include those for which the DGA recommends an upper bound (total fat, saturated fat, sodium, and cholesterol) and nutrients in the healthful category include those for which the DGA recommends a lower bound (fiber, iron, calcium, Vitamin A, and Vitamin C).

This index measure of diet quality has been commonly used in the public health literature (see, for example, Drewnowski, 2005, 2010). Each of the ratios $\left(\frac{pc_{jht}}{pc_j^{DGA}} \right)$ for nutrient j represents the Nutritional Quality Index (NQI) for that nutrient. The NQIs in and of themselves are informative about diet quality, and NQI values > 1 may be desirable or not depending on whether the nutrient is deemed healthful or not. The ratio of the sum of healthful and unhealthful NQIs then provides a comprehensive index measure of diet quality, where a higher nutrient ratio implies a higher quality diet. The nutrient ratio is a useful summary measure although it puts weight on micro nutrients as well as macro nutrients, making it less reflective of dietary advice.

All three of these measures are associated with whether someone in the household is diabetic, is obese, or has been diagnosed with hypertension (see Appendix Table A2) suggesting that they reflect consequential information about the quality of diet.

Summary statistics on these outcomes appear in Panel B of Table 1.

Data Limitations There are some limitations to the HomeScan data. The most important of these is that we observe only a subset of what households buy and consume. This occurs for two reasons: Nielsen does not include food away from home, and even within the subset of food at home it is likely not all purchases are recorded. Einav et al. (2010) provide validation and suggest that approximately half of trips are not recorded in HomeScan, although those which are recorded are highly accurate. Oster (2017) compares the HomeScan to the NHANES and to a benchmark calorie intake amount and suggests 65 to 80% of calories are recorded.

In this paper we primarily focus on outcomes based on diet *shares*. To the extent that we observe a random subset of purchases, then these shares will be an unbiased measure. Even if we do not see a random subset, if the treatment does not change scanning behavior we will have a measure of the impact. Issues will arise if, for example, a treatment changes scanning behavior differentially across food groups, or changes the consumption behavior with foods consumed away from home differently than foods consumed at home. This

[3]The recommendations can be found here: https://www.fda.gov/Food/IngredientsPackagingLabeling/LabelingNutrition/ucm274593.htm

is a limitation we will be unable to address.

2.2 Information Events

We merge the HomeScan data with other sources to perform our primary analysis of the impact of the information events of interest on diet quality.

Disease Diagnosis Data on disease diagnosis is drawn from the Nielsen Ailment Panel. This is a complementary survey in which some Nielsen panelists are surveyed about their health status. Panelists review a long list of diseases and indicate whether they have each one and, if yes, when they were diagnosed. The diagnosis measures are coarse - they indicate if it was in the last year, one to two years ago, three to four years ago or more than four years ago. We know the timing of the survey so we are able to use this to code diagnosis at the yearly level. In particular, the survey was run in January 2010. We identify households as newly diagnosed in 2009 if they report a diagnosis within the last year. We focus on three metabolic disease categories: (1) diabetes; (2) hypertension, high cholesterol and heart disease; and (3) obesity.

Table 2 shows summary statistics on disease prevalence and new diagnosis. The ailment data is available for 67,467 of the Nielsen households.

Government Programming The US Government issues general guidelines on food consumption. For many years, these guidelines were in the form of a pyramid, with food group shares indicated by either horizontal or vertical bars. In 2011, there was a change in messaging: the pyramid was turned into a plate (called "MyPlate"). The change in design was intended to make it easier for people to evaluate visually whether their diet is appropriate, and to emphasize the importance of fruits and vegetables in diet. The MyPlate visual has half of the plate taken up with the fruit and vegetable category. The initial messaging about the change to the plate was also focused around increasing fruit and vegetable consumption.

Research Findings Finally, we include as variables several major research findings on diet. Over the period of our data, three major studies were released about the Mediterranean diet: one in the *New England Journal of Medicine* (NEJM) in 2008, one in the *Journal of the American Medical Association* (JAMA) in 2009 and another in NEJM in 2013.[4] All three studies found large health benefits from this type of diet. Although the Mediterranean diet is defined as a whole diet, media coverage of the studies emphasized the role of olive oil and nuts. We therefore focus on olive oil and nut purchases as the outcome. All three studies - and especially the 2013 NEJM study - were widely covered in the media and resulted in spikes in internet interest in the topic (as measured by Google Trends).

[4]Shai et al, 2008; Féart et al, 2009; Estruch et al, 2013.

3 Baseline Results: Impacts of Information Events on Diet

This first section of results analyzes the impact of these treatments on average, and explores some baseline heterogeneity in impacts across demographics. We begin by describing the empirical strategy, and then discuss the results.

3.1 Empirical Strategy

Our first set of results focuses on disease diagnosis, an event which occurs at the household-year level. Our empirical strategy here uses a household fixed effects regression. The estimating equation is

$$y_{it} = \gamma_i + \tau_t + \beta \mathbf{T_{it}} + \epsilon_{it} \tag{1}$$

where y_{it} is the outcome (i.e. the measure of diet) for household i in year t, T_{it} is the vector of the treatment variable (years from disease diagnosis), γ_i is a household fixed effect and τ_t is a year fixed effect. The coefficient vector of interest is β. This regression is identified off of variation within a household over time. Although all diagnoses we consider occur sometime in 2009, the data includes non-diagnosed households, which allows us to separately identify treatment and year effects.

Our second set of results looks at changes in government dietary recommendations and research findings. These treatments vary only at the time level. We will therefore focus in on dietary changes right around the time of the event. Further, we will focus in each case on the set of foods which are targeted by the advice - fruits and vegetables in the case of MyPlate, and nuts and olive oil in the case of the Mediterranean diet studies. We will use other surrounding calendar years to adjust the data for calendar week controls, with the goal of addressing concerns about seasonality. The overall regression can be expressed as

$$y_{itw} = \gamma_i + \tau_w + \lambda_y + \beta \mathbf{T_{tw}} + \mu_{itw} \tag{2}$$

where y_{it} is the consumption of household i in calendar week w at time t. T_{tw} is the vector of the treatment variable (time relative to publication), γ_i is a household fixed effect, τ_w is a calendar week fixed effect and λ_y denotes a year fixed effect.

3.2 Results

Main Results Figure 1 shows the movements in one of our outcome variables - the share of unhealthy foods in total food spending - around disease diagnosis. Figure 1a shows the results around a diagnosis of hypertension, Figure 1b around a diagnosis of obesity and Figure 1c around a diagnosis of diabetes. In the case of diabetes, we see a reduction in the share of purchases of unhealthy foods around this event. This effect, while significant, is small: about 0.1 standard deviations. For hypertension and obesity there is no movement in the period after diagnosis. These figures focus on one of our outcomes, but similar effects can

be seen in Appendix Figures A1 and A2, which replicate Figure 1 for the healthy food expenditure share and the nutrient ratio.

Table 3 shows these effects statistically. There is again some suggestion of very small dietary improvements - a lower unhealthy food expenditure share, a higher healthy food expenditure share, a better nutrient ratio - after a diabetes diagnosis. There is no consistent change in behavior after diagnosis with hypertension or obesity. We include indicators of significance based on a standard set of p-value thresholds, and also note significance with a Bonferroni correction which adjusts for multiple hypothesis testing.

In general, the response is quite muted, despite the seemingly substantial treatment. One explanation for this is that the treated group is a sample which is, at least to some extent, selected to be resistant to behavior change. A diagnosis of diabetes, for example, typically comes after a period of warnings and failed behavior change. This group may be especially non-responsive (Oster, 2017).

To look for a behavioral response among a more general population, we turn to the impacts of government educational policies and research findings. On the one hand, these are much less notable events for any given individual. On the other hand, the targets of these approaches may include people with greater likelihood of response.

In order to facilitate identification we focus on changes in the immediate time vicinity of the news event. This allows us to be more confident in drawing causal conclusions if there is an effect, although it picks up only short term changes.

Figure 2 shows graphical evidence on the movements in the diet outcomes around the events. Panel A shows the evidence on fruit and vegetable purchases around the MyPlate announcement. Panel B shows the average diet study response. To improve precision, since the outcomes are the same, we aggregate the three diet studies together. Panel A shows no movement in fruit and vegetable purchases. Panel B shows a small increase in olive oil and nuts - this is significant in the first weeks, but fades quickly.[5]

Table 4 shows these results statistically. The format of this table mimics the format of Table 3 but with a shorter time frame - we are looking now in terms of weeks or months rather than years. The results mimic the evidence in Figure 2. There is some initial increase in olive oil and nuts, but as a share of a standard deviation it is very tiny (0.01 standard deviations). This is in spite of clear spikes in interest in these studies, as evidenced by Google Trends search behavior.

Heterogeneity On average, the observed changes in behavior are small. This may mask heterogeneity across demographic groups. A simple human capital theory would predict greater responsiveness among richer and more educated households, and among younger people. In addition, we expect some mechanical relationship between behavioral response and baseline levels, since people with (for example) more unhealthy purchases at baseline have more scope for improvement.

Table 5 estimates, for the unhealthy food share outcome, heterogeneity in response to disease diagnosis

[5]If we separate the three research studies we see the largest impact for the 2013 NEJM study but it is not as precise, and is still very small in magnitude.

across terciles of income, education, age and the baseline unhealthy share. We simplify the full timing analysis from above, and define a dummy for "after" - the first two years after diagnosis - versus "before" - one to two years before diagnosis. We exclude the year of diagnosis (2009). The first row in each panel of Table 5 shows the average effect, and the subsequent rows divide the sample by income, age and education. Appendix Tables A3 and A4 replicate the form of this table for the healthy food share and nutrient ratio, with similar patterns.

Overall, this table shows fairly little. There is some limited heterogeneity across demographics for diabetics, but this goes in the opposite direction of what one would predict: more highly educated household respond less. We do see is that those diabetics with a worse diet at baseline respond more, which may simply reflect that there are more ways for them to reduce their unhealthy food share. For the other diagnosis categories we do not see any notable results.

Table 6 shows a similar heterogeneity analysis for the response to policy and research findings. Again, the variation across groups is not systematic in a way that is consistent with human capital theory. Moreover, there are no groups with very large response. We do see a role for the baseline purchases here again, likely reflecting a mechanical effect.

The results in this section suggest little response to any of the variables considered, either on average or among likely target demographics. There are multiple interpretations of this finding. One is that the healthfulness of diet is largely fixed and unresponsive to information. A second interpretation is that this analysis does not successfully identify factors which influence behavior change, or does not successfully identify households who are likely to be more successful at changing behavior. These interpretations differ in their policy implications.

To evaluate the potential of the second interpretation, the following sections take an alternative approach. We begin by identifying a set of households which show sustained improvements in diet over an extended period. We then use a machine learning approach to identify the factors which predict this successful change. This includes a rich set of household and diet characteristics The model will allow us to identify whether change is predictable and, if so, what features, or combinations thereof, predict this change.

4 Identifying Successful Diet Changers

Our goal is to identify any households who successfully improve their diet by an amount that is both large enough to represent a meaningful change in diet quality and is sustained. We will refer to these household as "successful changers." Note that it is possible that the data does not contain any households of this type.

Identifying these changers with our data is empirically challenging. If we saw a perfectly complete measure of diet over many years, the exercise would be straightforward. In our data, however, we see a subset of what households consume and do not see an infinite length panel.[6] This leads to concerns about mean reversion.

[6]There is an additional issue here that for virtually all households diets are better in the first half of the year and worse in the second, and in general spending levels are correlated with diet shares, perhaps because people are more reliable about

To give a concrete example: one way to approach this would be to find households with a large improvement in diet quality from any one month to the next and define these as successful changer households. In practice, this will not yield a good measure of what we want. Households who show a large change after a month of low-quality diet are likely to be those with a generally good diet who just had a single outlying month.

Broadly, the problem is mean reversion. It is closely related to the discussion in Chay, McEwan and Uquiola (2005) (among others). In this case it is exacerbated by the fact that we do not see all purchases and people may vary in the fidelity with which they scan items over time.

We therefore need to define a dietary change in a way that limits this mean reversion problem. We do this in two ways. First, we limit our data to a sub-sample of households who are observed consistently over a long time period (30 months) and without major outlier months in terms of total spending. Specifically, we drop the bottom 5% of households in terms of total spending, and also any households with two or more months in a row with a total spending more than 2.5 standard deviations away from the average spending. Effectively, we try to begin with a sample where we are as close as possible to seeing a full measure of grocery purchases. This sample contains 45,987 households, versus 158,792 in the total sample.

Second, we identify household dietary changes based on a long pre-period and a long post-period. We identify two groups of successful changer households, one based on the share of unhealthy foods and the other based on the nutrient ratio measure. In either case, we require households to have an improvement (reduction in unhealthy foods or increase in nutrient ratio) over a ten month period, following a ten month baseline period.

Formally, define the diet quality measure in month t as H_t. Further, adopt the notation min_2 to indicate the second smallest value of a set, and max_2 as the second largest value of a set.

Given this, we define a month t as a "changer" month with respect to unhealthy foods if

$$min_2 \{H_{t-10}, H_{t-9}, ..., H_{t-1}\} \geq max_2 \{H_t, H_{t+1}, ..., H_{t+9}\} + c$$

where $c = 0.025$.

We define a month t as a "changer" month with respect to the nutrient ratio if

$$max_2 \{H_{t-10}, H_{t-9}, ..., H_{t-1}\} \geq min_2 \{H_t, H_{t+1}, ..., H_{t+9}\} + c$$

where $c = 0.06$. This requires, for example, the household to substantially reduce the share of unhealthy expenditures - by 2.5 percentage points - over an extended period, allowing for one deviating month in the pre- and post-period.

scanning some items than others. We adjust for this by initially residualizing all diet outcomes with respect to calendar month, total spending, and spending as a share of household average in each month.

Using the unhealthy share, this approach identifies 2,403 households with dietary improvement; this is 5.2% of the sample; using the nutrient ratio, there are 3,010 households, or 6.5% of the sample. For either outcome, the remainder of the (non-changer) households will comprise our comparison group. We assign them a random "change" date for the purposes of graphical comparison.

Figure 3a shows the evidence on unhealthy expenditure shares for the changer and non-changer households over a period of 40 months.[7] The twenty months in the middle of the graph are the period based on which the households are chosen. During this period there is a large reduction in the unhealthy food share; 15 percentage points on average.

The difference is smaller if we compare the pre-pre-periods (months -20 to -10) to the post-post-period (months 10 to 20) but there is still a large difference in unhealthy food share in the changer group before and after, versus no difference in the non-changer group. This suggests we are identifying changes which persist - at least to some extent - over time. This change - about 6 percentage points, or 25% of the mean - is large relative to other features of the data. For example, this is roughly twice as large as the cross-sectional difference in this variable between households with high school and college education.

Figure 3b shows the same graph for the identified nutrient ratio changers. Again, there is a large difference and it persists if we look outside the identification period.

The patterns in the data suggest that in the case of the unhealthy expenditure share the changer households have significantly worse diets in the pre-period, and then improve to have a diet comparable to the average in the longer run. In the case of the nutrient ratio, the changer population is similar to the rest *ex ante* and then improves to have a better diet in the post period. This suggests that we are picking up slightly different populations with the two definitions, although both groups show substantial dietary improvements.

Appendix Figures A3a and A3b show changes in the share of healthy foods in the purchase set and the nutrient ratio for the households selected based on unhealthy share. Similarly, Appendix Figures A4a and A4b show changes in the share of healthy foods and unhealthy foods for those households selected based on nutrient ratio. These other measures of diet also improve.

Table 7 shows basic summary statistics - including a summary measure of pre-period diet - for the households who change and those who do not, with selection based on the unhealthy share. The changer households are less well educated and are poorer than those who do not change. Also notable is that the baseline level of the outcome variable is higher for the changers than the non-changers, which is obvious from the figures.

Despite the differences in demographics in terms of t-tests, there is almost no aggregate predictive power of demographic characteristics in a simple regression. We regress changer status on baseline demographics and estimate the out of sample R-squared. It is close to zero: 0.0049 for the unhealthy changers; 0.0054 for the nutrient ratio changers.

Adding simple measures of baseline outcome measure to this regression improves the predictive power

[7]Note that although the middle 20 months are fully balanced, since we require only 30 months to be in the sample, the overall graph is slightly unbalanced.

somewhat, as would be suggested by the graph. However, the out of sample R-squared is still low: 0.07 for the unhealthy changers, and 0.007 for the nutrient ratio changers.

We turn now to whether it is possible to predict which households are likely to change using a learning model, and ask whether we can learn about the patterns of change from that evidence.

5 Predicting Dietary Changes

5.1 Random Forest Learning

Our goal is to predict which households will successfully improve their diet using a machine learning algorithm. Specifically, we use a random forest approach. The random forest is in the class of tree-based machine-learning prediction algorithms.[8] They have been used elsewhere in economics, although not frequently (e.g. Oster, 2017; Kleinberg et al, 2017), and are widely used in other fields. A key advantage of a random forest relative to some other machine learning algorithms, such as a lasso, is that it allows for non-linearities and automatic detection of interactions. This enables us to pick up potentially more complex patterns driving dietary change.

Broadly, the random forest starts from a prediction tree, which uses a set of inputs to predict an outcome. Tree-based methods work by partitioning units (here, households) based on their features into groups which are as similar as possible on the outcome (in this case, changing their behavior). The procedure works by generating a series of binary splits in the data based on the values of the input features. In the end, one is left with groups of households who are as similar as possible on the outcome, and share all the feature splits. These are the "leaves" of the tree.

Building only a single tree risks over-fitting. The random forest generates predictions by drawing many trees using bootstrapped samples of the data, and evaluating fit based on the out-of-sample performance of the prediction. Random forest is not the only approach to combining trees but it has been shown to perform well in a variety of applications.

The key input to the random forest is the feature set used in the prediction. In this case, we include in the feature set a rich set of demographics, baseline diet characteristics, and general dummies for calendar months and years. The diet characteristics include expenditures by category, as well as information on diet concentration and previous shorter lived dietary improvements.

We grow a random forest using 600 trees. We implement the random forest in R using the randomForest and randomForestSRC packages [9]. Further details on implementation are in Appendix B.

[8] We describe this briefly here, but interested readers can find more details about machine learning in general in Friedman, Hastie and Tibshirani (2009) and about random forests in particular in Breiman (2001).

[9] We implement the main random forest using the randomForest package. We then use the randomforestSRC package to detect interactions using a "minimal depth and maximal v-subtree" algorithm. See Appendix B for more details.

5.2 Results

A visual sense of the predictive power of the random forest output can be obtained by graphing the true positive rate versus the false positive rate. This captures the false positive rate that you have to accept to get a given true positive rate. A curve which lies on the 45 degree line has no predictive power: to get 50% of the true positives you have to admit 50% of the false positives. A curve which lies far above the line indicate more predictive power. The area under the curve (AUC) summarizes the strength of the prediction. An AUC of 0.5 is not predictive at all, and an AUC of 1 is perfectly predictive.

Figure 4a shows this curve for our prediction using the unhealthy changer sample; Figure 4b shows the same curve for the nutrient ratio changer sample. The AUCs are 0.81 and 0.71, respectively, which is moderately to highly predictive. Another way to see the predictive value is to say if we targeted the people with predictions in the top 10%, approximately 55% of them would be successful changers in the unhealthy case, versus 5% of the overall population. These figures are 37% and 6% for the nutrient ratio changer sample.

A key question for us is which features in the forest are predictive. A standard way to summarize this is by reporting variable importance. This identifies the most important features, where importance is defined as appearing in a large share of the trees and appearing at a higher tree split. Column 1 of Table 8 lists the top 10 features ranked by their importance. Panel A reports on the features predicting the unhealthy changers, and Panel B on features predicting the nutrient ratio.

In both cases the baseline level of the outcome is the most highly predictive feature. Other highly predictive features which differ across the outcomes are shares in various food groups. Perhaps most interesting is to look for features which are common across the panels and which do not relate to the baseline levels. In both cases, measures of diet concentration - the standard deviation of the shares across groups - are highly predictive. This suggests that behavioral improvements are related to the variability of the diet *ex ante.* Diet concentration is high for households where a few food groups account for a large share of purchases.

From our standpoint, the standard importance measure misses several things. In particular, we are not solely (or even largely) interested in the quality of the final prediction here. Instead, we are interested in whether this identifies some particular features or combination of a few features which are associated with behavior change. To comment on this we need, at a minimum, to understand the shape of the relationship between behavior change and the important features. The importance ranking tells us only what is important in prediction, not the structure of the relationship. Beyond this, one of the main characteristics of the random forest is that it allows for interactions between variables. Some of these interactions play a more important role than others, and their structure may be informative about patterns driving change. However, the importance plot summarizes only single variable importance.

To further develop these results, therefore, we adopt the visualization approach developed in Hastie et al. (2009) and Jones and Linder (2015). In brief, this approach allows us to make partial dependence plots illustrating non-parametrically the relationship between a feature x and the outcome (in this case behavior

change). The illustrated relationship includes the averaged effects of all the interactions of x with the other features, which allows it to capture the dependence relationship we see in the data.[10] Two-dimensional relationships can be represented by showing these partial dependence plots for some x by groups of another variable. More details are in Appendix B.

We begin by visualizing the relationship between single variables and the outcome. Figure 5 focuses on the unhealthy food changer sample. Panel (a) shows the relationship with baseline level of unhealthy foods and Panels (b) and (c) with two measures of diet concentration. Panel (a) shows a clear upward sloping relationship. Households with a worse diet at baseline are more likely to improve their diet. This relationship is not surprising and has, to some extent, a mechanical interpretation. Those who eat more unhealthy food to start have more to reduce.

Panels (b) and (c) show that diet concentration also has a positive relationship with behavior change. For the metric using TFP in particular, there is a strong upward sloping relationship between concentration and subsequent behavior change.

Figure 6 shows a similar figure for the nutrient ratio changer sample. Here, Panel (a) shows the relationship with the baseline nutrient ratio, and Panels (b) and (c) illustrates diet concentration. Panel (a) in particular illustrates some of the value of this visualization approach. The relationship between the baseline nutrient ratio and the outcome is U-shaped: change is most likely for people with either high or low nutrient ratio at baseline. This is not captured in the importance plot.

Panels (b) and (c) show that the relationship between change and dietary concentration looks very similar to the relationship in the unhealthy change group. For both concentration measures, it slopes significantly upward.

These visualization tools can also be used to helpfully identify and illustrate interactions between variables in the random forest. As a first step, Column 2 of Table 8 lists the top ten interactions, ranked by importance. Again, Panel A is the unhealthy changer group and Panel B is the nutrient ratio group. Many of the important interactions in both cases include the variables which are important on their own, interacted with each other. Again, without visualization this can be a bit opaque.

Figure 7 illustrates one key interaction for the unhealthy changers - that between the diet concentration and baseline unhealthy share. Panel (a) uses the TFP concentration measure, Panel (b) the Nielsen group concentration measure. Change is most common among households that have high values on both features. That is, those whose diet is concentrated and who have a poor diet at the start. The conclusion is the same in both panels.

The predictive power of these elements in combination is substantial. The random forest suggests a chance of behavior change of 40 to 50% among those households with the highest concentration and the largest baseline share. The base rate is 5%, so targeting these households would (in expectation) identify a considerably more responsive set.

[10] This partial dependence plot is distinct from a marginal dependence plot, which would represent the marginal impact holding other features constant. Instead, this will capture the fact that other features correlate with x.

Figures 8a and 8b show the same interactions for nutrient ratio. These are less stark, although in Figure 8b in particular we can see evidence that the relationship with concentration is stronger for those with a worse nutrient ratio at baseline.

Taken together, this evidence suggests that diet concentration is a consistent predictor of dietary improvement on either metric. Baseline diet also plays a role, and may interact with concentration, particularly when we focus on changes in the unhealthy food share. This suggests substantial improvements in diet are more likely for households with a worse diet *ex ante*, those whose diets are more concentrated, and especially households where those two features interact.

5.3 Diet Concentration and Disease Diagnosis

Before concluding, we turn briefly to applying the findings from the random forest to our initial analysis of dietary response to disease diagnosis from Section 3. We focus on the disease diagnosis case.

A key result in Section 3 is that there is little predictable heterogeneity across demographic lines. The random forest results confirm these findings, and suggest other variables - specifically, dietary concentration - which may play a more important role in heterogeneity.

Table 9 replicates the form of Table 5, but focusing on the role of diet concentration. This table focuses on changes in the unhealthy share, as in Table 5 (other outcomes appear in the appendix tables).

The results here show more consistent evidence of heterogeneity across concentration. Focusing in particular on the group with the largest changes - diabetics - households with more concentrated diets change their behavior more. Diabetics in the highest terciles of diet concentration reduce their unhealthy food share by about 3 percentage points. This is a sizable change, equal to about 15 percent of the mean and roughly as large as the cross-sectional gap between college and non-college educated households.

In the third row of each panel we show the impact of diet concentration among only those households in the top tercile of unhealthy share at baseline. This is intended to capture the role of the interaction between the baseline diet quality and diet concentration. Indeed, for diabetics we see that the most concentrated households in this row have almost a 4 percentage point decrease in unhealthy share. In this case, for both hypertension and obesity we begin to see some reduction – around 1 percentage point – for those households with a concentrated diet and a high baseline unhealthy share, though these changes do not reach statistical significance.

In the end, this analysis suggests that the diet concentration - identified by the random forest as a key predictor of change in the population overall - also predicts responsiveness to disease diagnosis. Particularly for diabetics, this suggests an alternative targeting approach.

6 Conclusion

This paper analyzes the determinants of changes in diet in a general population. We find, first, that even in response to seemingly large treatments - for example, major disease diagnosis - changes in diet are very limited. Research findings and changes in government policy advice seem to have a similarly negligible role. There is little predictable heterogeneity.

We find, however, that there are a small share of households who do show large improvements in diet over time. Using a machine learning model we look to predict who these households are. We find that some components of baseline diet are successful predictors of behavior change. Notably, households with a concentrated diet *ex ante* are more likely to improve their diet. We show that this dietary concentration is also predictive of change after disease diagnosis.

We argue for two broad conclusions. First, the fact that behavior change is so limited even after such large events suggests that it may be a challenge to use education to change behavior. This is consistent with a literature in economics and elsewhere showing that diets are not especially malleable (e.g. Atkin, 2016). Second, the finding from the random forest suggests that there may be overlooked components of diet which could predict change. Machine learning, particularly when combined with the visualization tools here, may provide a way to develop these insights.

References

Allcott, Hunt, Rebecca Diamond, and Jean-Pierre Dube, "The geography of poverty and nutrition: Food deserts and food choices across the United States," Technical Report, National Bureau of Economic Research 2017.

Atkin, David, "The caloric costs of culture: Evidence from Indian migrants," *The American Economic Review*, 2016, *106* (4), 1144–1181.

Bleich, Sara N, David Cutler, Christopher Murray, and Alyce Adams, "Why is the developed world obese?," *Annu. Rev. Public Health*, 2008, *29*, 273–295.

Breiman, Leo, "Random forests," *Machine learning*, 2001, *45* (1), 5–32.

Broadbent, E., L. Donkin, and J. C. Stroh, "Illness and treatment perceptions are associated with adherence to medications, diet, and exercise in diabetic patients," *Diabetes Care*, Feb 2011, *34* (2), 338–340.

Bronnenberg, Bart J, Jean-Pierre H Dube, and Matthew Gentzkow, "The evolution of brand preferences: Evidence from consumer migration," *The American Economic Review*, 2012, *102* (6), 2472–2508.

Carrera, Mariana, Syeda A Hasan, and Silvia Prina, "The Effects of Health Risk Assessments on Cafeteria Purchases: Do New Information and Health Training Matter?," 2017.

Chay, Kenneth Y, Patrick J McEwan, and Miguel Urquiola, "The central role of noise in evaluating interventions that use test scores to rank schools," *The American Economic Review*, 2005, *95* (4), 1237–1258.

Cutler, David M, Edward L Glaeser, and Jesse M Shapiro, "Why have Americans become more obese?," *The Journal of Economic Perspectives*, 2003, *17* (3), 93–118.

Delamater, Alan M., "Improving Patient Adherence," *Clinical Diabetes*, 2006, *24* (2), 71–77.

Drewnowski, Adam, "Concept of a nutritious food: toward a nutrient density score," *The American journal of clinical nutrition*, 2005, *82* (4), 721–732.

___ , "The Nutrient Rich Foods Index helps to identify healthy, affordable foods," *The American journal of clinical nutrition*, 2010, *91* (4), 1095S–1101S.

Dubois, Pierre, Rachel Griffith, and Aviv Nevo, "Do Prices and Attributes Explain International Differences in Food Purchases?," *American Economic Review*, March 2014, *104* (3), 832–67.

Einav, Liran, Ephraim Leibtag, and Aviv Nevo, "Recording discrepancies in Nielsen Homescan data: Are they present and do they matter?," *Quantitative Marketing and Economics*, 2010.

Estruch, R., E. Ros, J. Salas-Salvado et al., "Primary prevention of cardiovascular disease with a Mediterranean diet," *N. Engl. J. Med.*, Apr 2013, *368* (14), 1279–1290.

Féart, Catherine, Cécilia Samieri, Virginie Rondeau, Hélène Amieva, Florence Portet, Jean-François Dartigues, Nikolaos Scarmeas, and Pascale Barberger-Gateau, "Adherence to a Mediterranean diet, cognitive decline, and risk of dementia," *Jama*, 2009, *302* (6), 638–648.

Feldstein, A. C., G. A. Nichols, D. H. Smith, V. J. Stevens, K. Bachman, A. G. Rosales, and N. Perrin, "Weight change in diabetes and glycemic and blood pressure control," *Diabetes Care*, Oct 2008, *31* (10), 1960–1965.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani, *The elements of statistical learning*, Vol. 2, Springer series in statistics Springer, Berlin, 2009.

Gilchrist, Duncan Sheppard and Emily Glassberg Sands, "Something to Talk About: Social Spillovers in Movie Consumption," *Journal of Political Economy*, 2016, *124* (5), 1339–1382.

Group, Diabetes Prevention Program Research et al., "10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study," *The Lancet*, 2009, *374* (9702), 1677–1686.

Handbury, Jessie, Ilya Rahkovsky, and Molly Schnell, "Is the Focus on Food Deserts Fruitless? Retail Access and Food Purchases Across the Socioeconomic Spectrum," *National Bureau of Economic Research*, 2015.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, "Overview of supervised learning," in "The elements of statistical learning," Springer, 2009, pp. 9–41.

Ishwaran, Hemant, Udaya B Kogalur, Eiran Z Gorodeski, Andy J Minn, and Michael S Lauer, "High-dimensional variable selection for survival data," *Journal of the American Statistical Association*, 2010, *105* (489), 205–217.

Jones, Zachary and Fridolin Linder, "Exploratory data analysis using random forests," in "Prepared for the 73rd annual MPSA conference" 2015.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human decisions and machine predictions," Technical Report, National Bureau of Economic Research 2017.

Oster, Emily, "Diabetes and Diet: Purchasing Behavior Change in Response to Health Information," *American Economic Journal: Applied Economics*, forthcoming.

Pi-Sunyer, Xavier, "The look AHEAD trial: a review and discussion of its outcomes," *Current nutrition reports*, 2014, *3* (4), 387–391.

Ponzo, V., R. Rosato, E. Tarsia, I. Goitre, F. De Michieli, M. Fadda, T. Monge, A. Pezzana, F. Broglio, and S. Bo, "Self-reported adherence to diet and preferences towards type of meal plan in patient with type 2 diabetes mellitus. A cross-sectional study," *Nutr Metab Cardiovasc Dis*, Jul 2017, *27* (7), 642–650.

Raj, G. D., Z. Hashemi, D. C. Soria Contreras, S. Babwik, D. Maxwell, R. C. Bell, and C. B. Chan, "Adherence to Diabetes Dietary Guidelines Assessed Using a Validated Questionnaire Predicts Glucose Control in Individuals with Type 2 Diabetes," *Can J Diabetes*, Jun 2017.

Shai, Iris, Dan Schwarzfuchs, Yaakov Henkin, Danit R Shahar, Shula Witkow, Ilana Greenberg, Rachel Golan, Drora Fraser, Arkady Bolotin, Hilel Vardi et al., "Weight loss with a low-carbohydrate, Mediterranean, or low-fat diet," *N Engl J Med*, 2008, *2008* (359), 229–241.

Swinburn, Boyd, Gary Sacks, and Eric Ravussin, "Increased food energy supply is more than sufficient to explain the US epidemic of obesity," *The American journal of clinical nutrition*, 2009, *90* (6), 1453–1456.

Volpe, Richard, Abigail Okrent, and Ephraim Leibtag, "The effect of supercenter-format stores on the healthfulness of consumers' grocery purchases," *American Journal of Agricultural Economics*, 2013, p. 132.

Table 1: **Summary Statistics**

	Mean	*Standard Deviation*	*Sample Size*
Panel A: Panelist Demographics			
HH Head Age	47.7	10.3	158,792
HH Head Years of Education	14.4	3.3	158,792
HH Income	$65,932	$45,690	158,792
HH Size	2.59	1.33	158,792
White (0/1)	0.82	0.38	158,792
Panel B: Panelist Shopping Behavior			
Nutrient Ratio	0.501	0.288	158,792
Average Duration in Panel (Months)	47.4	39.5	158,792
Shopping Behavior:			
Calories (person/month)	28,286	14,385	158,726
Expenditures (person/month)	$81.48	$41.79	158,742
Expenditure Shares on:			
Whole Grains	2.04%	1.97%	158,787
Non-Whole Grains	18.0%	5.7%	158,787
Potato Products	1.69%	1.24%	158,787
Dark Green Vegetables	1.55%	1.64%	158,787
Orange Vegetables	0.44%	0.56%	158,787
Beans, Lentils, Peas	0.31%	0.47%	158,787
Other Vegetables	2.85%	2.17%	158,787
Whole Fruits	3.35%	3.03%	158,787
Fruit Juices	1.92%	1.35%	158,787
Whole Milk Products	3.94%	3.49%	158,787
Low Fat/Skim Milk Products	3.47%	3.58%	158,787
All Cheese	5.24%	3.11%	158,787
Beef, Pork, Veal, Lamb, Game	5.64%	3.62%	158,787
Chicken, Turkey	0.68%	1.10%	158,787
Fish	1.34%	1.67%	158,787
Bacon, Sausage, Lunch Meats	1.39%	1.71%	158,787
Nut, Nut Butters, Seeds	2.29%	2.24%	158,787
Egg and Egg Mixtures	1.26%	1.17%	158,787
Fats, Condiments	2.15%	1.57%	158,787
Coffee, Tea	2.23%	2.68%	158,787
Soft Drinks, Soda	5.61%	5.42%	158,787
Sugars, Sweets, Candy	18.1%	7.1%	158,787
Soups	5.61%	2.64%	158,787
Frozen Entrees	8.90%	6.02%	158,787

Note: This table shows summary statistics for the Nielsen Panelists. Household age, income, and education are computed at the median of reported categories. Mean values for calories and expenditures are generated after a 1% winsorization.

Table 2: **Disease Diagnosis Information**

	Total # Households	*Any Diagnosis*	*New Diagnosis, 2009*
Hypertension/High Cholesterol/Heart Disease	67,467	52.1%	5.5%
Obesity	67,467	22.0%	1.9%
Diabetes	67,467	11.7%	0.8%

Note: This table shows summary statistics on disease diagnosis for individuals surveyed in the Nielsen Ailment Panel. The survey was run in early 2010. Diagnosis in 2009 is inferred from reporting a new diagnosis in the last year.

Table 3: **Response to Diagnosis**

	(1) Yr -3 b/se/sd	(2) Yr -2 b/se/sd	(3) Yr -1 Mean b/se/sd	(4) Treat Yr b/se/sd	(5) Yr +1 b/se/sd	(6) Yr +2 b/se/sd
A. Unhealthy Share:						
Hypertension, cholesterol, heart	0.002498	0.000947	0.237798	-0.001357	-0.002294	-0.001749
	(0.00219)	(0.00164)		(0.00167)	(0.00193)	(0.00204)
	[0.018]	[0.007]		[-0.010]	[-0.017]	[-0.013]
Obesity	-0.004207	0.000367	0.240478	0.000660	-0.000588	-0.001332
	(0.00451)	(0.00330)		(0.00336)	(0.00423)	(0.00413)
	[-0.031]	[0.003]		[0.005]	[-0.004]	[-0.010]
Diabetes	-0.000205	0.003834	0.244416	-0.011508***	-0.016635***+	-0.013733**
	(0.00601)	(0.00385)		(0.00441)	(0.00504)	(0.00555)
	[-0.001]	[0.028]		[-0.084]	[-0.121]	[-0.100]
B. Healthy Share:						
Hypertension, cholesterol, heart	-0.000982	-0.001772*	0.111350	-0.000287	0.002971**	0.001478
	(0.00142)	(0.00106)		(0.00104)	(0.00134)	(0.00142)
	[-0.011]	[-0.020]		[-0.003]	[0.033]	[0.016]
Obesity	-0.003502	-0.004139	0.113620	-0.001136	-0.000381	-0.003078
	(0.00361)	(0.00277)		(0.00227)	(0.00319)	(0.00328)
	[-0.039]	[-0.046]		[-0.013]	[-0.004]	[-0.034]
Diabetes	0.000190	0.002349	0.107718	0.003756	0.009357***+	0.001508
	(0.00377)	(0.00297)		(0.00310)	(0.00320)	(0.00344)
	[0.002]	[0.026]		[0.042]	[0.104]	[0.017]
C. Nutrient Ratio:						
Hypertension, cholesterol, heart	-0.001643	-0.002664	0.506821	0.004326	0.016003***+	0.008990*
	(0.00508)	(0.00386)		(0.00370)	(0.00440)	(0.00471)
	[-0.006]	[-0.009]		[0.015]	[0.055]	[0.031]
Obesity	-0.002789	0.009464	0.497952	0.001898	0.005390	0.001025
	(0.01144)	(0.00926)		(0.00790)	(0.00936)	(0.01034)
	[-0.010]	[0.033]		[0.007]	[0.019]	[0.004]
Diabetes	-0.018023	-0.012052	0.474394	0.015990*	0.025707**	-0.005254
	(0.01248)	(0.00922)		(0.00968)	(0.01107)	(0.01013)
	[-0.062]	[-0.042]		[0.055]	[0.089]	[-0.018]

Standard errors in parentheses, standard deviation change in square brackets.

P-values significance: * = 0.1, ** = 0.05, *** = 0.01, ***+ = significant at 0.05 using a Bonferroni correction.

Notes: This table shows the impact of disease diagnosis on food purchases. The "Unhealthy Share" is the share of expenditure on unhealthy food categories (soft drinks, soda, fruit drinks and ades, sugar, sweets and candy). The "Healthy Share" is the share of expenditure on healthy food categories (whole grains, fruits, vegetables). Details on the Nutrient Ratio construction are in Section 2.

Table 4: **Response to Government Policy, Research**

	(1) Mnth -2 b/se/sd	(2) Mnth -1 b/se/sd	(3) Wk -1 Mean b/se/sd	(4) Publ Wk b/se/sd	(5) Wk +1 b/se/sd	(6) Mnth +1 b/se/sd	(7) Mnth +2 b/se/sd
A. MyPlate 2011							
Fruits/Vegetables	0.001082	0.000342	0.103733	0.001677*	-0.000658	-0.000531	-0.000484
	(0.00079)	(0.00087)		(0.00102)	(0.00102)	(0.00082)	(0.00080)
	[0.008]	[0.003]		[0.013]	[-0.005]	[-0.004]	[-0.004]
B. Medit Diet Research							
Oil/Nuts	0.000081	0.000256	0.017222	0.000525**	0.000702***+	0.000630***+	0.000332*
	(0.00020)	(0.00022)		(0.00025)	(0.00026)	(0.00020)	(0.00020)
	[0.001]	[0.004]		[0.008]	[0.010]	[0.009]	[0.005]
Standard errors in parentheses, standard deviation change in square brackets.							
P-values significance: * = 0.1, ** = 0.05, *** = 0.01, ***+ = significant at 0.05 using a Bonferroni correction.							

Notes: This table shows the impact of government information policy and research findings on purchases. The outcome in Panel A is purchases of all fruits and vegetables, which was the focus of the initial messaging around MyPlate. The outcome in Panel B is purchase of olive oil and nuts, which are emphasized as key components of the Mediterranean diet. These regressions combine three studies (NEJM 2008; JAMA 2009; NEJM 2013) which find substantial impacts of the Mediterranean diet on health.

Table 5: **Heterogeneity in Unhealthy Purchase Share Response to Diagnosis**

	(1)	(2)	(3)	(4)
	Avg	Tercile 1	Tercile 2	Tercile 3
A. Hypertension, Cholesterol, Heart	-0.00171			
	(0.00186)			
Income		-0.00523	0.000199	0.000422
		(0.00341)	(0.00314)	(0.00309)
Age		-0.00240	0.000278	-0.00298
		(0.00353)	(0.00320)	(0.00296)
Education		-0.00607	-0.00254	0.000696
		(0.00398)	(0.00325)	(0.00274)
Baseline unhealthy food share		-0.00140	0.000367	-0.00294
		(0.00216)	(0.00241)	(0.00319)
B. Obesity	-0.00212			
	(0.00406)			
Income		0.000268	0.00170	-0.00815
		(0.00744)	(0.00730)	(0.00617)
Age		-0.00163	0.00104	-0.00639
		(0.00659)	(0.00724)	(0.00676)
Education		-0.00458	0.00474	-0.00478
		(0.00773)	(0.00653)	(0.00616)
Baseline unhealthy food share		-0.00424	0.00910	-0.00454
		(0.00482)	(0.00615)	(0.00507)
C. Diabetes	-0.0192***			
	(0.00495)			
Income		-0.0269***	-0.0276***	-0.00247
		(0.00800)	(0.00859)	(0.00882)
Age		-0.00484	-0.0248***	-0.0209***
		(0.00855)	(0.00890)	(0.00742)
Education		-0.0296***	-0.0206**	-0.0136*
		(0.0113)	(0.00830)	(0.00716)
Baseline unhealthy food share		-0.00530	-0.0140**	-0.0299***
		(0.00546)	(0.00645)	(0.00684)

Standard errors in parentheses

Education terciles: high school, some college, college/grad.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the heterogeneity in the impact of disease diagnosis by demographic groups. The outcome is the share of expenditure on unhealthy food categories (soft drinks, soda, fruit drinks and ades, sugar sweets and candy). Parallel tables for healthy food purchases and the nutrient ratio appear in Appendix Tables A3 and A4.

Table 6: **Heterogeneity in Response to Policy and Research**

	(1)	(2)	(3)	(4)
	Avg	Tercile 1	Tercile 2	Tercile 3
A. MyPlate 2011	-0.000416			
	(0.000529)			
Income		-0.000656	0.000288	-0.00232***
		(0.000626)	(0.000589)	(0.000873)
Age		-0.000567	-0.000265	-0.000442
		(0.000628)	(0.000638)	(0.000659)
Education		-0.00125*	-0.000719	0.0000513
		(0.000725)	(0.000653)	(0.000595)
Baseline Fruit+Veg		0.00373***	0.000858	-0.00549***
		(0.000563)	(0.000609)	(0.000734)
B. Mediterranean Diet Studies	0.00169***			
	(0.000510)			
Income		0.00181***	0.00180***	0.00144***
		(0.000541)	(0.000540)	(0.000558)
Age		0.00137**	0.00148***	0.00208***
		(0.000542)	(0.000544)	(0.000545)
Education		0.00168***	0.00162***	0.00174***
		(0.000568)	(0.000547)	(0.000531)
Baseline Oil+Nuts		0.00690***	0.00478***	-0.00501***
		(0.000509)	(0.000509)	(0.000591)

Standard errors in parentheses

Education tertiles: high school, some college, college/grad.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the heterogeneity in the impact of policy and research findings by demographic groups. The outcome in Panel A is purchases of all fruits and vegetables, which was the focus of the initial messaging around MyPlate. The outcome in Panels B is purchase of olive oil and nuts, which are emphasized as key components of the Mediterranean diet. These regressions combine three studies (NEJM 2008; JAMA 2009; NEJM 2013) which find substantial impacts of the Mediterranean diet on health.

Table 7: **Demographics of Large Change Households**

	Unhealthy Share Changers			Nutrient Ratio Changers		
	No Change	Change	p-value, diff	No Change	Change	p-value, diff
Age:						
Household Head Age	55.61	55.93	.2158	55.5	55.85	.1405
Income:						
HH Income: < 50K	.4811	.5393	2.5e-08	.486	.4679	.05426
HH Income: 50K to 100 K	.3837	.3598	.01829	.381	.3684	.1667
HH Income: > 100K	.1352	.1009	1.4e-06	.1329	.1636	1.7e-06
Education:						
High School or Less	.1946	.2174	.005886	.1979	.1684	.000083
Some College	.3001	.3211	.02919	.2985	.2786	.02038
Completed College	.3374	.3269	.288	.3396	.3502	.2379
Post College/Grad	.1679	.1346	.000019	.164	.2028	3.1e-08
Composition:						
Household size	2.363	2.088	1.7e-26	2.366	2.12	4.2e-26
Ethnicity:						
White	.8529	.8604	.3127	.8512	.8527	.8304
N	43584	2403		42782	3010	

Notes: This table compares demographics for the set of households identified as having a large dietary change in our data. Large changer households are those who show a sustained period of improved diet. In the first three columns the change is defined by a decline in the share of unhealthy foods in the diet of at least 2.5 percentage points. In the second set of columns the change is defined by an increase in the nutrient ratio of at least 0.06.

Table 8: **Important Features and Interactions in Random Forest**

(1) Top Importance Single Features	(2) Top Importance Interactions
Panel A: Unhealthy Share Changer Prediction	
Unhealthy share	Unhealthy share & SD TFP
SD TFP	Unhealthy share & SD unhealthy groups
Max TFP	Unhealthy share & carbonated beverages
SD unhealthy groups	Unhealthy share & max TFP
Ice cream	Unhealthy share & whole grains
Whole grains	Unhealthy share & ice cream
Carbonated beverages	Unhealthy share & SD Nielsen groups
SD Nielsen groups	Unhealthy share & soups
Candy	SD TFP & Max TFP
Soups	SD TFP & SD unhealthy groups
Panel B: Nutrient Ratio Changer Prediction	
Nutrient ratio	Nutrient ratio & fresh produce
Fresh produce	Nutrient ratio & potatoes
Potato	Fresh produce & potatoes
Packaged meats	Nutrient ratio & condiments, gravies, sauces
Baking supplies	Nutrient ratio & SD TFP
Condiments, gravies, sauces	Nutrient ratio & fats and condiments
SD Nielsen groups	Fresh produce & SD TFP
SD TFP	Nutrient ratio & soft drinks carbonated
Fats and condiments	Nutrient ratio & SD groups

Notes: This table shows the top importance features for the random forest, both the levels and the top interactions. Panel A focuses on predicting an improvement in the unhealthy share, Panel B on predicting improvement in the nutrient ratio. All food groups are measured in terms of share of total food spending.

Table 9: **Heterogeneity in Unhealthy Food Share Response by Diet Concentration**

	(1)	(2)	(3)	(4)
	Avg	Tercile 1	Tercile 2	Tercile 3
A. Hypertension, Cholesterol, Heart	-0.00166			
	(0.00155)			
SD TFP		-0.00213	-0.00535**	0.00170
		(0.00210)	(0.00249)	(0.00352)
SD Nielsen Groups		-0.000338	-0.00298	-0.00214
		(0.00208)	(0.00242)	(0.00372)
SD Nielsen Groups within Unhealthy share T3		-0.000651	0.000110	-0.00848
		(0.00480)	(0.00474)	(0.00659)
B. Obesity	-0.000395			
	(0.00331)			
SD TFP		-0.00240	0.00191	-0.00295
		(0.00458)	(0.00506)	(0.00716)
SD Nielsen Groups		0.000977	-0.00485	-0.0000575
		(0.00429)	(0.00468)	(0.00870)
SD Nielsen Groups within Unhealthy share T3		0.00103	-0.000196	-0.0157
		(0.00767)	(0.00682)	(0.0110)
C. Diabetes	-0.0172***			
	(0.00390)			
SD TFP		-0.0134***	-0.0201***	-0.0180**
		(0.00474)	(0.00633)	(0.00855)
SD Nielsen Groups		-0.00917**	-0.0162***	-0.0287***
		(0.00464)	(0.00608)	(0.00954)
SD Nielsen Groups within Unhealthy share T3		-0.0161*	-0.0293***	-0.0373**
		(0.00949)	(0.00754)	(0.0151)

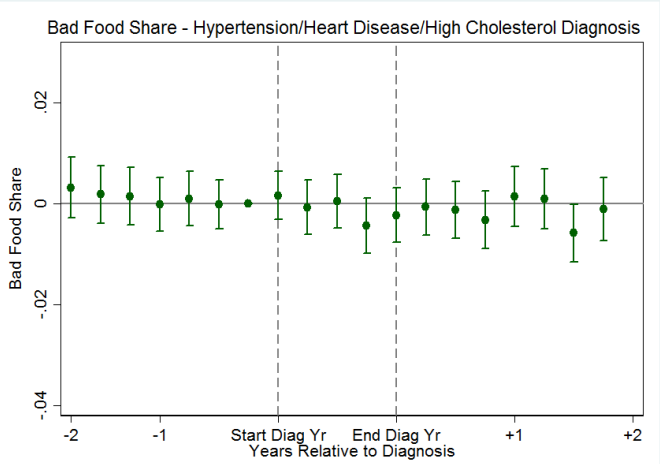
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

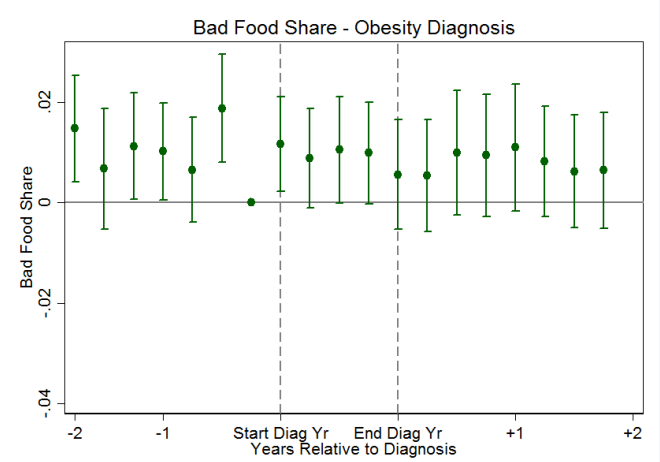
Notes: This table shows heterogeneity in behavioral response across baseline diet concentration. This concentration is measured as either the standard deviation of diet shares across TFP groups, or across Nielsen product groups.

Figure 1: **Effects of Diagnosis on Diet**

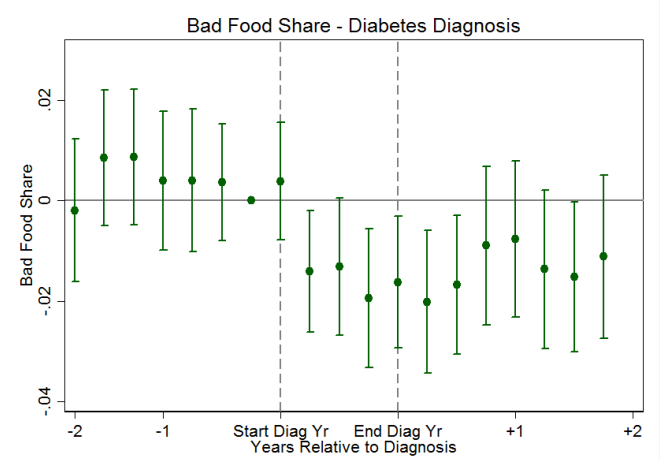
Panel A: Hypertension Diagnosis



Panel B: Obesity Diagnosis



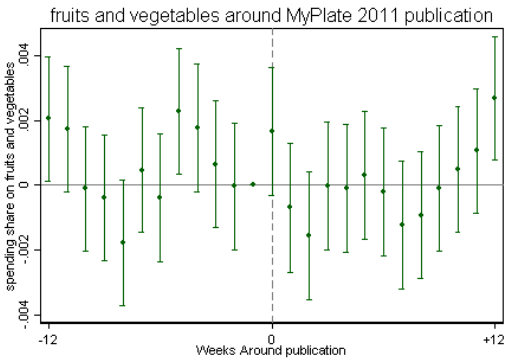
Panel C: Diabetes Diagnosis



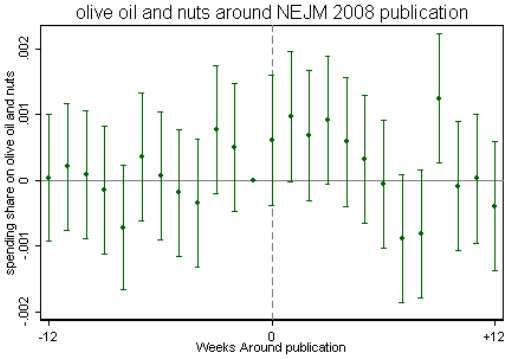
Notes: This figure shows the effect of diagnosis on purchase behavior for three diseases. The purchase behavior considered here is the share of expenditure in unhealthy foods (soft drinks, soda, fruit drinks and ades, sugar, sweets and candy). Parallel figures for healthy food purchases and nutrient ratio appear in the Appendix Figures A1 and A2. The coefficients are derived from the regression specified in Equation (1). The diagnosis year refers to the year during which the person reports diagnosis; we do not see more detailed timing than this.

Figure 2: **Response to Policy, Diet Studies**

Panel A: My Plate Study



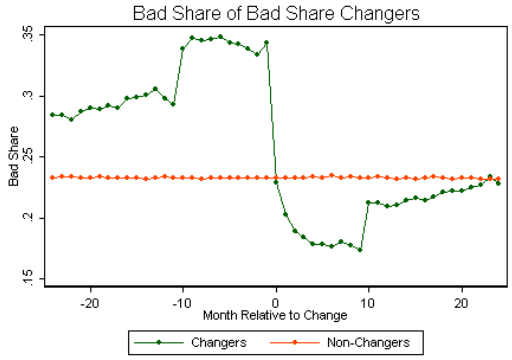
Panel B: Research Findings on Mediterranean Diet



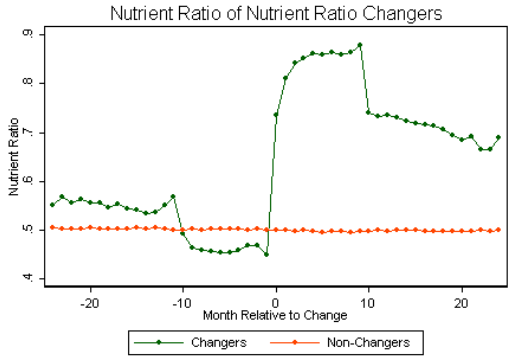
Notes: This figure shows the effect of government information policy (Panel A) and research findings (Panel B) on dietary choices. The outcome in Panel A is purchases of all fruits and vegetables, which was the focus of the initial messaging around MyPlate. The outcome in Panels B is purchase of olive oil and nuts, which are emphasized as key components of the Mediterranean diet. All three studies included find substantial health benefits of this type of diet.

Figure 3: **Changes in Diet Share for Identified Changer Households**

Panel A: Unhealthy Share Changers



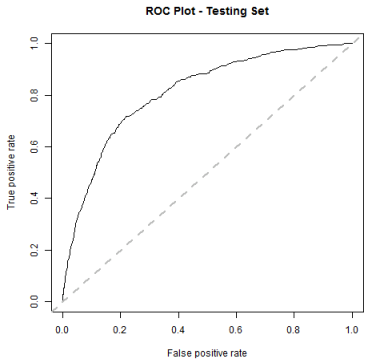
Panel B: Nutrient Ratio Changers



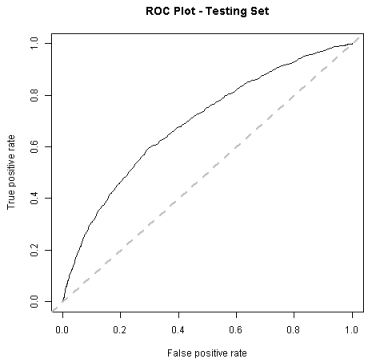
Notes: This figure shows the trend in the share of expenditures on unhealthy foods for households identified as "large changers" and those who are not. In Panel A the large changers are identified by having a reduction in the unhealthy expenditure share of at least 2.5 percentage points over the 20 months surrounding the change period. In Panel B the large changers are identified by having an increase in the nutrient ratio of at least 0.06 over the 20 months surrounding the change period.

Figure 4: **Random Forest Output: Prediction Quality**

Panel A: Unhealthy Share Changers



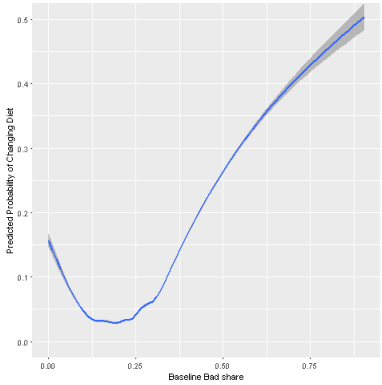
Panel B: Nutrient Ratio Changers



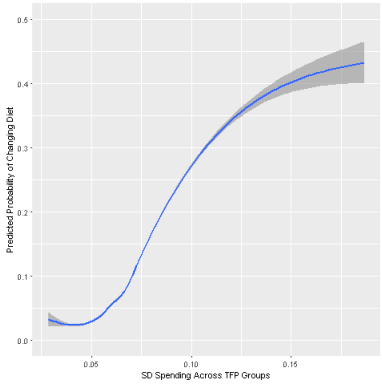
Notes: This figure shows the ROC curves from the random forest algorithm. In Panel A the large changers are identified by having a reduction in the unhealthy expenditure share of at least 2.5 percentage points over the 20 months surrounding the change period. In Panel B the large changers are identified by having an increase in the nutrient ratio of at least 0.06 over the 20 months surrounding the change period. The full list of features used in the random forest appears in Appendix B.

Figure 5: **Partial Dependence Plots for Unhealthy Changers**

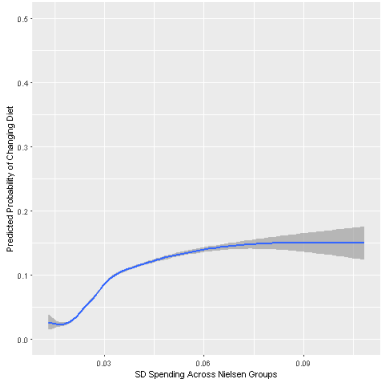
Panel A: Role of Baseline Unhealthy Share



Panel B: Role of Baseline Diet Concentration (TFP Groups)



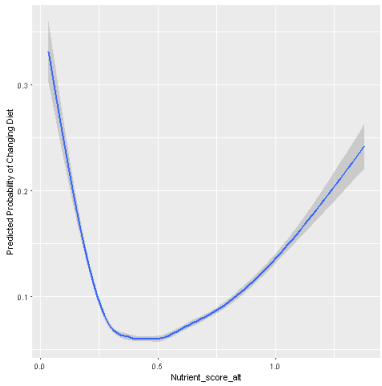
Panel C: Role of Baseline Diet Concentration (Nielsen Groups)



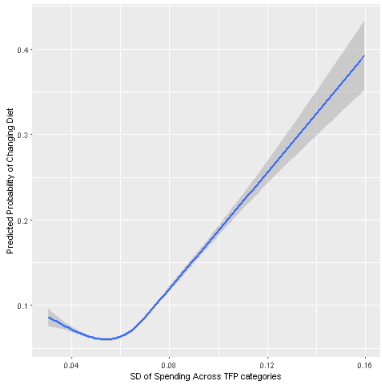
Notes: This figure shows the partial dependence plots for the prediction of change along the unhealthy food share dimension. These plots are generated as described in Jones and Linder (2015), see Appendix B. The plots capture the partial dependence between each feature and the outcome, averaging over the characteristics which appear in the data alongside that feature.

Figure 6: **Partial Dependence Plots for Nutrient Ratio Changers**

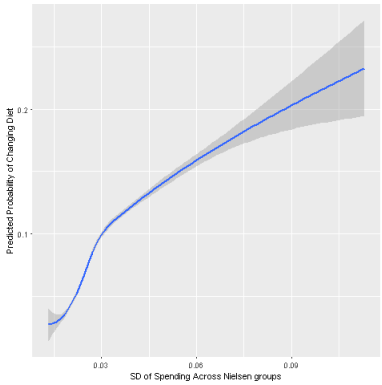
Panel A: Role of Baseline Nutrient Ratio



Panel B: Role of Baseline Diet Concentration (TFP Groups)



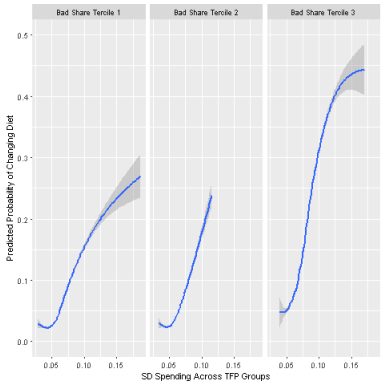
Panel C: Role of Baseline Diet Concentration (Nielsen Groups)



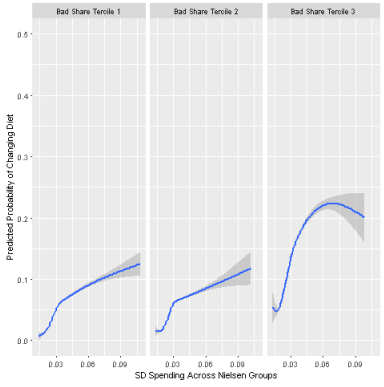
Notes: This figure shows the partial dependence plots for the prediction of change along the nutrient ratio dimension. These plots are generated as described in Jones and Linder (2015), see Appendix B. The plots capture the partial dependence between each feature and the outcome, averaging over the characteristics which appear in the data alongside that feature.

Figure 7: **Interaction Plots for Unhealthy Changers**

Panel A: Baseline Share and TFP Concentration



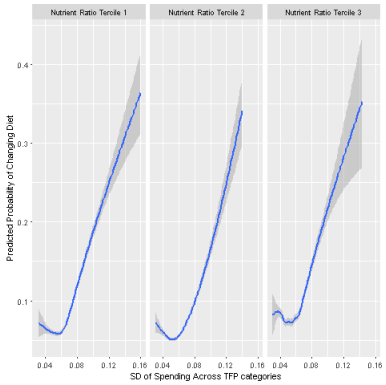
Panel B: Baseline Share and Nielsen Group Concentration



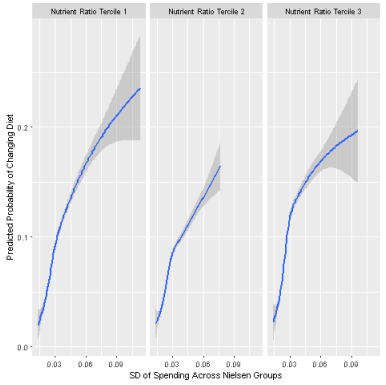
Notes: This figure shows the interacted partial dependence plots for the prediction of change along the unhealthy food share dimension. These plots are generated as described in Jones and Linder (2015), see Appendix B. The plots capture the partial dependence between diet concentration and the outcome, averaging over the characteristics which appear in the data alongside that feature. The relationships are further disaggregated across terciles of the baseline unhealthy share.

Figure 8: **Interaction Plots for Nutrient Ratio Changers**

Panel A: Baseline Nutrient Ratio and TFP Concentration



Panel B: Baseline Nutrient Ratio and Nielsen Group Concentration



Notes: This figure shows the interacted partial dependence plots for the prediction of change along the nutrient ratio dimension. These plots are generated as described in Jones and Linder (2015), see Appendix B. The plots capture the partial dependence between diet concentration and the outcome, averaging over the characteristics which appear in the data alongside that feature. The relationships are further dis-aggregated across terciles of the baseline nutrient ratio.

Appendix A: Tables and Figures

Table A1: **Comparison of Demographics: Nielsen versus US Census**

	Nielsen Mean	US Census Mean
HH Head Age	47.4 (10.3)	48.9
HH Head Years of Education	14.4 (3.3)	13.7
HH Income	65,932 (45,690)	68,918
HH Size	2.59 (1.33)	2.58
White (0/1)	0.82 (0.38)	0.72

Note: This table shows the demographics of the Nielsen sample relative to the US census in 2010. Standard errors in parentheses.
** Indicates difference from the census mean at the 5% level.

Table A2: **Relationship between Diet and Health Outcomes**

	Hypertension/Cholesterol/Heart			Obesity			Diabetes		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Nutrient ratio	-0.00444***			-0.0130***			-0.0100***		
	(-16.91)			(-57.62)			(-55.68)		
Unhealthy share		0.00764***			0.00150***			0.00385***	
		(29.04)			(6.66)			(21.50)	
Healthy share			0.00745***			-0.00614***			-0.000890***
			(29.07)			(-27.99)			(-5.09)
Observations	3474231	3645638	3645638	3474231	3645638	3645638	3474231	3645638	3645638
R^2	0.000	0.000	0.000	0.001	0.000	0.000	0.001	0.000	0.000

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, ***$p < 0.001$

Note: This table shows the relationship between our measures of diet and disease. Each coefficient is interpretable as the impact of a one standard deviation increase in the diet measure on the chance of a diagnosis in the household.

Table A3: **Heterogeneity in Healthy Food Share Response to Diagnosis**

	(1)	(2)	(3)	(4)
	Avg	Tercile 1	Tercile 2	Tercile 3
A. Hypertension, Cholesterol, Heart	0.00198*			
	(0.00102)			
Income		0.00413**	-0.000845	0.00301
		(0.00182)	(0.00158)	(0.00187)
Age		0.000573	0.00170	0.00351**
		(0.00188)	(0.00178)	(0.00164)
Education		0.00436**	-0.00172	0.00362**
		(0.00208)	(0.00165)	(0.00159)
Baseline healthy food share		-0.0000115	-0.0000207	0.00553***
		(0.00159)	(0.00152)	(0.00206)
B. Obesity	0.000673			
	(0.00237)			
Income		0.00108	0.000921	0.00143
		(0.00362)	(0.00448)	(0.00417)
Age		0.00331	0.00135	-0.00382
		(0.00374)	(0.00360)	(0.00523)
Education		0.00423	-0.000757	0.00120
		(0.00410)	(0.00387)	(0.00362)
Baseline healthy food share		-0.00201	0.00472	0.000545
		(0.00319)	(0.00420)	(0.00433)
C. Diabetes	0.00490*			
	(0.00251)			
Income		0.00428	0.00574	0.00604
		(0.00341)	(0.00490)	(0.00480)
Age		0.00101	0.00771*	0.00543
		(0.00574)	(0.00429)	(0.00367)
Education		0.00546	0.00714	0.00427
		(0.00354)	(0.00446)	(0.00418)
Baseline healthy food share		0.0102***	0.00409	-0.00210
		(0.00369)	(0.00350)	(0.00566)

Standard errors in parentheses

Education Terciles: high school, some college, college/grad.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the heterogeneity in the impact of disease diagnosis by demographic groups. The outcome is the share of expenditure on healthy food categories (fruits, vegetables, whole grains).

Table A4: **Heterogeneity in Nutrient Score Response to Diagnosis**

	(1)	(2)	(3)	(4)
	Avg	Tercile 1	Tercile 2	Tercile 3
A. Hypertension, Cholesterol, Heart	0.0100***			
	(0.00358)			
Income		0.0104*	0.00316	0.0172***
		(0.00613)	(0.00625)	(0.00632)
Age		0.00810	0.0149***	0.00671
		(0.00672)	(0.00575)	(0.00627)
Education		0.0139*	0.00111	0.0146***
		(0.00774)	(0.00578)	(0.00553)
Baseline nutrient ratio		0.00838*	0.00965*	0.0119
		(0.00499)	(0.00518)	(0.00758)
B. Obesity	0.000740			
	(0.00732)			
Income		-0.00542	0.00274	0.00657
		(0.0121)	(0.0125)	(0.0135)
Age		0.00930	-0.0105	0.00284
		(0.0127)	(0.0111)	(0.0149)
Education		0.00557	-0.00501	0.00314
		(0.0171)	(0.0123)	(0.0106)
Baseline nutrient ratio		0.0212**	-0.0143	-0.0139
		(0.00902)	(0.0116)	(0.0148)
C. Diabetes	0.0242***			
	(0.00878)			
Income		0.0379**	0.00737	0.0288*
		(0.0151)	(0.0151)	(0.0155)
Age		0.0264	0.0138	0.0358***
		(0.0185)	(0.0162)	(0.0122)
Education		0.0169	0.0314*	0.0255**
		(0.0153)	(0.0185)	(0.0126)
Baseline nutrient ratio		0.0136	0.0122	0.0332
		(0.0111)	(0.0144)	(0.0227)

Standard errors in parentheses

Education Terciles: high school, some college, college/grad.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows the heterogeneity in the impact of disease diagnosis by demographic groups. The outcome is the nutrient score of the overall purchase basket.

Table A5: **Heterogeneity in Healthy Food Share Response by Diet Concentration**

	(1)	(2)	(3)	(4)
	Avg	Tertile 1	Tertile 2	Tertile 3
A. Hypertension, Cholesterol, Heart	0.00198*			
	(0.00102)			
SD TFP		0.00382**	0.00105	0.00172
		(0.00166)	(0.00172)	(0.00195)
SD Nielsen Groups		0.00174	0.00166	0.00303
		(0.00144)	(0.00162)	(0.00233)
SD Nielsen Groups within Healthy share T1		0.00251	-0.00533***	0.00348
		(0.00252)	(0.00207)	(0.00312)
B. Obesity	0.000673			
	(0.00237)			
SD TFP		0.000437	0.00413	-0.00224
		(0.00413)	(0.00392)	(0.00412)
SD Nielsen Groups		-0.00113	0.0000640	0.00528
		(0.00266)	(0.00362)	(0.00642)
SD Nielsen Groups within Healthy share T1		0.000489	-0.00125	-0.00424
		(0.00439)	(0.00441)	(0.00669)
C. Diabetes	0.00490*			
	(0.00251)			
SD TFP		0.00667	0.00486	0.00481
		(0.00515)	(0.00387)	(0.00384)
SD Nielsen Groups		0.00653*	-0.000520	0.0110**
		(0.00397)	(0.00385)	(0.00514)
SD Nielsen Groups within Healthy share T1		0.00211	0.0128**	0.0121**
		(0.00699)	(0.00578)	(0.00594)

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Notes: This table shows heterogeneity in behavioral response across baseline diet concentration. This concentration is measured as either the standard deviation of diet shares across TFP groups, or across Nielsen product groups.

Table A6: **Heterogeneity in Nutrient Ratio Response by Diet Concentration**

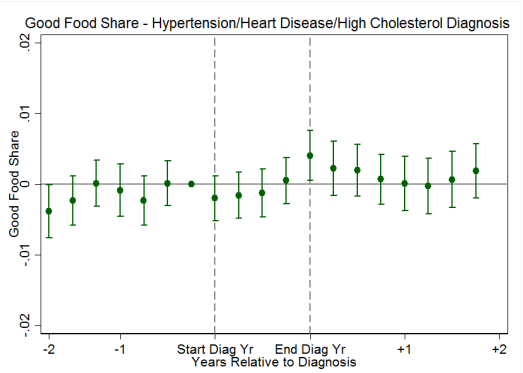
	(1)	(2)	(3)	(4)
	Avg	Tertile 1	Tertile 2	Tertile 3
A. Hypertension, Cholesterol, Heart	0.0100***			
	(0.00358)			
SD TFP		0.0140**	0.00809	0.00955
		(0.00619)	(0.00573)	(0.00693)
SD Nielsen Groups		0.0138***	-0.00436	0.0239***
		(0.00525)	(0.00612)	(0.00762)
SD Nielsen Groups within Nutrient ratio T1		0.0167**	0.00360	0.00526
		(0.00752)	(0.00922)	(0.00896)
B. Obesity	0.000740			
	(0.00732)			
SD TFP		-0.0000717	0.0202*	-0.0210*
		(0.0130)	(0.0121)	(0.0127)
SD Nielsen Groups		-0.00410	0.0169	-0.0119
		(0.0112)	(0.0121)	(0.0152)
SD Nielsen Groups within Nutrient ratio T1		0.0264	0.0279**	0.00596
		(0.0173)	(0.0112)	(0.0177)
C. Diabetes	0.0242***			
	(0.00878)			
SD TFP		0.0485***	0.00139	0.0250
		(0.0164)	(0.0117)	(0.0169)
SD Nielsen Groups		0.0230*	0.0153	0.0375*
		(0.0124)	(0.0136)	(0.0205)
SD Nielsen Groups within Nutrient ratio T1		0.0233	-0.0231*	0.0345
		(0.0168)	(0.0120)	(0.0222)

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

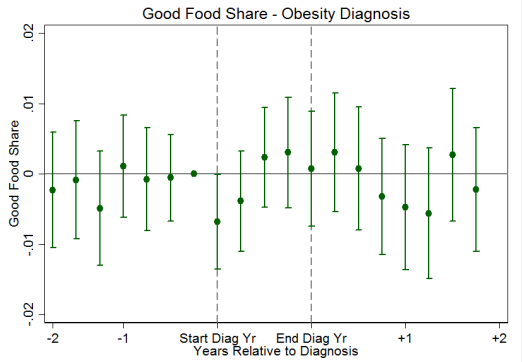
Notes: This table shows heterogeneity in behavioral response across baseline diet concentration. This concentration is measured as either the standard deviation of diet shares across TFP groups, or across Nielsen product groups.

Figure A1: **Disease Diagnosis and Healthy Food Share**

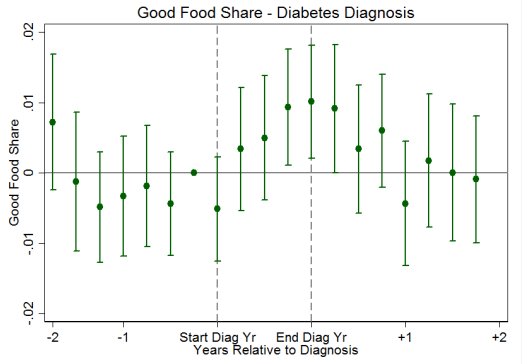
Panel A: Healthy Food Share: Hypertension



Panel B: Healthy Food Share: Obesity



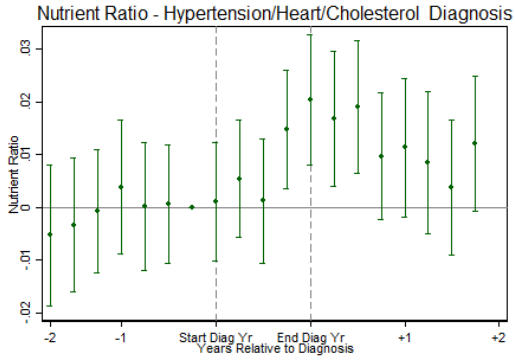
Panel C: Healthy Food Share: Diabetes



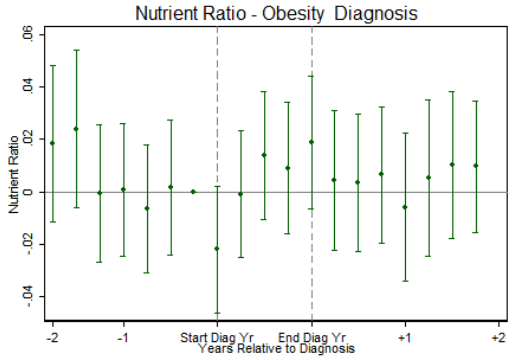
Notes: This figure shows the effect of diagnosis on purchase behavior for three diseases. The purchase behavior considered here is the share of expenditure in healthy foods (vegetables, fruits, and whole grains). The coefficients are derived from the regression specified in Equation (1). The diagnosis year refers to the year over which the person reports diagnosis; we do not see more detailed timing than this.

Figure A2: **Disease Diagnosis and Nutrient Ratio**

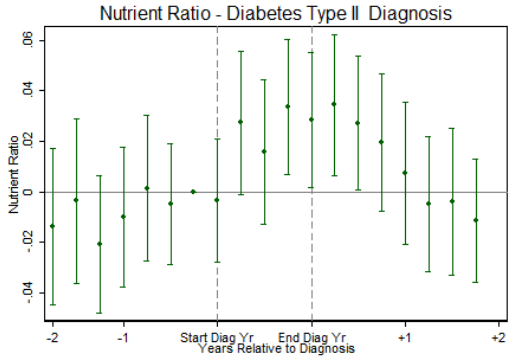
Panel A: Nutrient Ratio: Hypertension



Panel B: Nutrient Ratio: Obesity



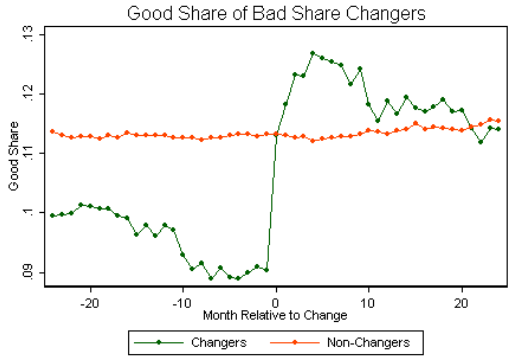
Panel C: Nutrient Ratio: Diabetes



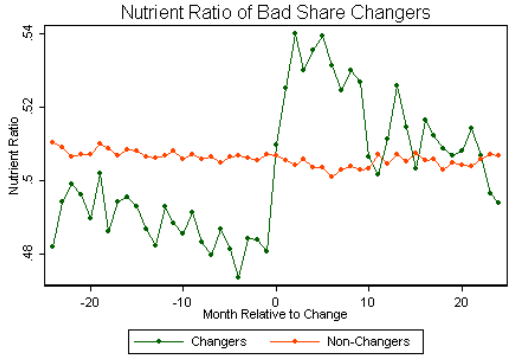
Notes: This figure shows the effect of diagnosis on purchase behavior for three diseases. The purchase behavior considered here is the nutrient ratio for the overall basket. The coefficients are derived from the regression specified in Equation (1). The diagnosis year refers to the year over which the person reports diagnosis; we do not see more detailed timing than this.

Figure A3: **Auxiliary Changes in Large Changers: Unhealthy Change Definition**

Panel A: Healthy Food Share



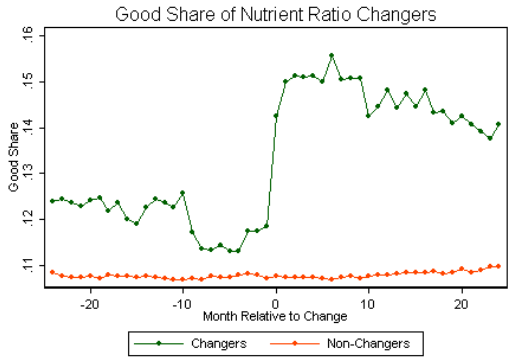
Panel B: Nutrient Ratio



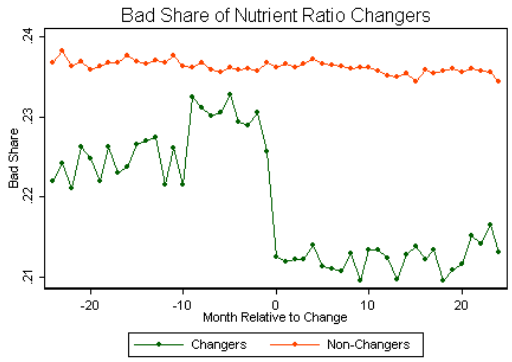
Notes: This figure shows the share of purchases in healthy foods (fruits, vegetables, whole grains) for household we define as successful diet changers versus non-changers in Panel A. Panel B shows the changes in nutrient ratio. Changer status is defined as a reduction in the share of purchases in unhealthy foods by at least 2.5 percentage points over a sustained period of 10 months after a 10 month baseline.

Figure A4: **Auxiliary Changes in Large Changers: Nutrient Ratio Change Definition**

Panel A: Good Food Share



Panel B: Unhealthy Food Share



Notes: This figure shows the share of purchases in healthy foods (fruits, vegetables, whole grains) for household we define as successful diet changers versus non-changers in Panel A. Panel B shows the changes in unhealthy food share. Changer status is defined as an increase in nurient ratio over a sustained period of 10 months after a 10 month baseline.

Appendix B: Machine Learning Details

B.1 Random Forest Implementation

The random forest was created in R using the **randomforest** package. The following features are included in the analysis:

- Household demographics (age group, education, income, household size, marital status, employment status, and race)
- Changes in other health behaviors, including magnitude of changes in alcohol and smoking expenditures, dummy variables for whether these changes were large, and dummy variables for whether the head of the household quit smoking
- Other health behaviors (amount of spending on cigarettes, alcohol, diet aids, and vitamins)
- Local health and economic characteristics (median household income and obesity rate)
- Share of expenditures in each food category as defined by the USDA Thrifty Food Plan (TFP)
- Share of expenditures on each food item as defined by the Nielsen product groups
- Concentration of food expenditures across all food categories (standard deviation and maximum of spending among TFP and Nielsen groups)
- Concentration of food expenditures across unhealthy food categories (standard deviation of spending among unhealthy TFP groups)
- Average fraction of food expenditures on unhealthy food categories
- Average difference between household nutrition in January and the rest of the year

The command is built with 600 trees in classification mode with a node size of 1. The output is predicted probability.

B.2 Interaction Identification and Partial Dependence Plots

Interaction detection was completed with the **randomForestSRC** package in R using a "maximal v-subtrees and minimal depth" algorithm as described in Jones and Linder (2015). The random forest was created with 1000 trees, using the same features as above. The algorithm then measures the importance of an interaction between two variables w and v by averaging the "minimal depth of w in the maximal subtree of v" (Jones and Linder 2015) for all of the trees in the random forest. A maximal subtree of w refers to the largest subtree that has a root node on w (see Ishwaran et al., 2010). The intuition behind the procedure is that features with a higher importance appear at higher splits (closer to the root node) in each tree. The minimal depth of v represents the distance between the highest maximal subtree for variable v and the root node of the whole tree, and is therefore a measure of variable importance. This idea can also be applied to detect interactions by examining so-called second-order maximal subtrees (Ishwaran et al., 2010). The interaction between v and w can be captured by calculating the minimal depth of w in the maximal subtree of v, and averaging this across the trees in the forest.

The output is an n-by-n matrix, where n is the number of variables in the random forest. The values on the diagonals represent the relative importance of individual variables, normalized between 0 and 1, and the values on the off-diagonals represent the importance of interactions as calculated by the "maximal v-subtrees and minimal depth" method. Comparing the relative values of the off-diagonal entries allows for ranking the importance of interactions between each pair of features.

Partial dependence plots are used to visualize the size and direction of the interactions identified above. The partial dependence plots are presented as a modification of the plots in Jones and Linder (2015). As the authors describe in more detail, partial dependence plots are created by generating a synthetic dataset for each value of the variable of interest. This value is assigned to all observations, while the other features in the data are left unchanged. Each synthetic dataset is then 'dropped' down the forest and used to generate predicted probabilities. Averaging over these predictions generates the mean predicted probability of change for each value of the variable of interest. These are graphically represented in a partial dependence plot.

To construct the partial dependence plots for two variables, we group one variableinto terciles of its values among observations in the random forest. Within each tercile, a curve is plotted to represent the impact of changes in the second variable on the predicted probability of the outcome variable from the random forest. A 95% confidence interval for the curve is also constructed at each value. Differences in the shapes of the curves across terciles visually indicate an interaction between the two variables in the random forest. Partial dependence plots were constructed using the random forests from the **randomForestSRC** package using the **ggplot** visualization package in R.