Changes in Household Diet: Determinants and Predictability *

Stefan Hut
Brown University

Emily Oster
Brown University and NBER

January 2, 2019

Abstract

We use grocery purchase data to analyze dietary changes. We show that households – including those with more income or education - do not improve diet in response to disease diagnosis or changes in household circumstances. We then identify households who show large improvements in diet quality. We use machine learning to predict these households and find (1) concentration of baseline diet in a small number of foods is a predictor of improvement and (2) dietary changes are concentrated in a small number of foods. We argue these patterns may be well fit by a model which incorporates attention costs.

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

*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

expensive for the health care system and have both morbidity and mortality consequences for individuals (Feldstein et al., 2008). Further, 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 quality of diet is therefore a social and policy goal.

This paper takes a data-driven approach to these questions. First, we document a series of facts - some confirming existing work, some new - on overall dietary change and on the features of households with successful diet improvements. Second, we interpret these facts through the lens of a model and argue this may speak to policy.

We begin with the facts.

There is a growing literature within economics which seeks to estimate what changes in information, circumstances or food options will prompt changes in diet behavior. These papers estimate the impact of changes in food type availability (Handbury et al, 2015; Bronnenberg et al., 2012; Allcott et al., 2017, Hut, 2018), food prices (Atkin, 2016) and health status (Oster, 2018). The general approach is to estimate behavioral response to events and, in most cases, explore demographic heterogeneity in response. This approach links with predictions of economic theory suggesting that changes in information or the relative cost of different choices should influence behavior, and the further implication that some groups (better educated, richer) should respond more than others.

These papers provide little evidence that dietary choices are malleable. The quality of diet seems on average insensitive to these stimuli, and in some cases people seem to be willing to give up substantial health benefits to maintain their preferred diet (Atkin, 2016; Oster, 2018).

The first part of this paper extends these null results, using the Nielsen HomeScan panel, a grocery scanner dataset. Nielsen participant households scan grocery (and other) purchases with a home hand scanner. We construct a measure of diet quality - a "diet score" - using information from a survey of doctors on the quality of food groups.[2] We confirm earlier findings that healthfulness of household diet is largely unresponsive to diabetes diagnosis

[2]This measure is described in more detail below and correlates with other measures of diet quality used in existing literature.

(Oster, 2018) and then show it also does not change with diagnosis of other metabolic disease (hypertension, heart disease, obesity). We show further that significant changes in household circumstances (childbirth, marriage, divorce, job loss, retirement, etc.) do not change dietary health. This lack of change is despite the fact that the average diet in the sample is quite poor.

We confirm that the lack of dietary response also holds for more highly educated, younger and richer households. Combined with the earlier work, this suggests that large dietary improvements are, at best, rare.

There are, however, some empirical examples in other data of significant dietary change. Feldstein et al (2008), for example, shows substantial weight loss among a small share of people diagnosed with diabetes. Identifying these individuals and learning about their features, as well as what dietary changes they actually make, may be key to targeting policy and to better understanding the limits on behavior change.

The second part of this paper, therefore, takes an alternative approach to the data. Rather than attempting to identify events which prompt behavior change, we use the data to find households who show sustained improvements in diet quality. We then ask to what extent these households can be predicted, what predicts successful change and what changes in diet are actually made. Our goal here is to generate an auxiliary set of facts - beyond the observation that behavior change is unusual - which can be used to better inform theory.

The Nielsen HomeScan panel data is the central input to the analysis. The structure and size of the data provide some substantial advantages. The panel is large; overall, we observe 160,000 households for an average of 48 months each. In addition, many households are in the panel for a long period, allowing a detailed look at diet over a long time. The large sample size is useful when we are looking for rare events like large behavior changes.[3]

In Section 4 of the paper we turn to the identification of households with substantial improvements in diet quality. A central challenge is the problem of mean reversion; this is a particular issue given that our data is not a perfect measure of all food consumption. To limit concerns about this we focus on a balanced panel of households with a sustained

[3]The data also has some limitations. 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. These are discussed in more detail in the data section.

period of change following a long baseline period. We identify approximately 7% of the sample as households with substantial improvements in diet quality. On average, these make large changes, about triple the cross-sectional difference between the score of high school and college graduates.

Having identified these households, we ask two questions. First, are behavior changes predictable, based either on events, demographics or baseline diet? Second, *how* do households change their behavior - what food groups, for example, seem most susceptible to change?

For the problem of prediction, we use a machine learning approach - in particular, a random forest algorithm. This allows us to estimate the predictive role of a wide variety of events and diet features without concerns about over-fitting. We find that the overall predictive power of events (including those we describe above, and others like changes in government diet recommendations, major research findings) is nil. Similarly, demographics - either baseline demographics or changes in household circumstances - also have little predictive power. Baseline dietary patterns, however, do appear to predict dietary improvement.

Baseline diet quality impacts subsequent change, and both households with better and worse diets are more likely to change. In addition, we identify dietary concentration as a key driver. Household with more concentrated diets - namely, diets where a large share of expenditures are in a smaller number of groups - are more likely to improve the quality of their diet in the subsequent months of the data.

Turning to the question of what changes these households actually make, we find that dietary changes are also concentrated. A small number of changes account for an outsize share of the overall dietary improvements among these households. For example, on average 60% of the diet score improvement is accounted for by the Nielsen product group with the largest change. Even when we limit to product modules - a much narrower classification such as cookies or yogurt - we find that the group with the largest change accounts for on average a third of the diet score change, and the top two modules account for half.

Following the empirical analysis, we turn to discussing implications of these results for theory of behavior change. We focus on explaining three facts from the data. First, substantial behavior change is rare and does not seem to be prompted by particular events. Second, it is more common for households with a concentrated diet to change their diet.

Third, when households do change their diet the changes are concentrated in a small number of food groups. These facts are difficult to explain in a simple neoclassical model. Although limited behavior change overall can be accommodated by a number of modifications, a neoclassical model will generally predict that when households do improve their diet, they will do so through small changes in many categories. This derives from a standard assumption of concave utility of consumption in any good.

We suggest that a model which includes attention costs (similar to Gabaix, 2014) may provide a better fit to the data. In this model the household has to pay a cost for each food group they change, which we theorize is the cost of "paying attention" to that category. We show that this produces the prediction that dietary changes are concentrated, and that households with more concentrated diets at baseline will have a lower cost of changing. Overall, the cost of changing diet in this model is larger than in a standard model, which may help explain the rarity of such changes.

From a policy standpoint, this observation may point to the value of policies which emphasize simple diet rules or the value of making a small number of changes, rather than making more blanket recommendations like "eat a balanced diet" or "consume fewer calories".

The first part of 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 et al., 2011; Ponzo et al, 2017; Raj et al, 2017) as well as 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 et al., 2017; Oster, 2018).

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, 2018; 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 is organized as follows. Section 2 describes the data we use, and Section 3 shows the first set of facts on responses to disease and household circumstance. Section 4 describes our identification of "large change" households, and Section 5 analyzes

the changes in this group. Section 6 discusses theory in light of the facts developed, and Section 7 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 in the panel are asked to scan their purchases after all shopping trips; this includes grocery and pharmacy purchases, large retailer and super-center purchases, as well as purchases made online and at smaller retailers. The Nielsen data records the Universal Product Code (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. 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 the racial composition is more heavily white.

Our focus is on the healthfulness of diet. To analyze this, we need to convert the purchase behavior to metrics of diet quality. The data contains UPC-level detail about the foods purchased, so in principle we can look in great detail at what people purchase. We are interested, however, in the general healthiness of the diet, not necessarily in individual foods. Two households may achieve a similarly healthy (or unhealthy) diet in different ways.

A natural approach to aggregation would be to combine nutrition data at the UPC level; indeed, other papers, including by these authors, does so (Oster, 2018; Hut, 2019; Alcott et al, 2017; Dubios, Griffin and Nevo, 2014). However, merging with nutrition data introduces considerable noise, as only slightly more than half of UPC codes can typically be matched directly, and therefore many nutrition facts must be imputed. The quality of these imputations varies over time. In addition, even given nutrition data, there are some questions of how to combine it into an aggregate quality measure.

As an alternative, we make use of a survey of doctors which focused on the quality of food groups in the HomeScan data. We implemented a survey of 17 primary care doctors, providing the survey participants with a list of food items which we could link to the HomeScan data (example: applesauce, lettuce, breakfast bars). We asked the doctors to indicate, in general, whether they thought this category was a "Good source of Calories", a "Bad Source of Calories" or "Neither good nor bad."

In 80% of cases all or all but one of the doctors agreed on the ranking of the food, suggesting we are picking up something consistent about the foods.

Based on the survey response, we assign each doctor-item observation a score p_{jn}, for food item j and doctor n. This score equals 1 if the item was considered good, 0 if considered neither good nor bad, or -1 if considered bad. We aggregate these scores across doctors to assign each Nielsen food module j a point score $p_j = \frac{\sum_n p_{jn}}{17}$, where $p_j \in [-1, 1]$. As such, $p_j = 1$ if all doctors consider the module good, and $p_j = -1$ if all doctors consider the module bad. A perfectly neutral module has a score of $p_j = 0$. We then create an overall diet score as a weighted average of these point values and module spending shares. Specifically, for individual i we have

$$score_i = \sum_j share_{ij} p_j$$

where $share_{ij}$ is the expenditure share of household $i's$ basket which is made up of food item j. The score is scaled between -1 and 1, so a household who consumes only foods which all doctors agree are bad would have a score of -1, and one who consumes only foods which all doctors agree are good would have a score of 1.

The average household has a diet score of -0.296 (see Panel B of Table 1). As some

background, Panel B of Table 1 also shows the shares of purchases in the five largest Nielsen product groups.

This measure of diet quality has the advantage of using expenditures directly and we argue it serves to capture information about the diet advice that doctors would give patients about how to change their diet. In practice, it is closely related to other metrics. In Appendix Table A2 we show this is highly correlated with a variety of other summary diet measures. This provides some comfort that our evidence is not unique to this metric of diet quality.

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 (*forthcoming*) compares the HomeScan to the NHANES and to a benchmark calorie intake amount and suggests 65 to 80% of calories are recorded.

Our primary outcome in this paper is based on the relative amounts of good and bad foods in the diet, not absolute levels of food spending. 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 is a limitation we will be unable to address.

2.2 Household Events

We merge the HomeScan data with other sources to look for events which may correlate with changes in diet.

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 (January 2010) so we are able to use this to code diagnosis at the yearly level. 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.

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

Changes in Household Circumstances Data on changes in household circumstances are drawn from the Nielsen panel information on demographics. Panelists are surveyed yearly about their demographic profiles, so we use this to construct information on changes over time.

Based on the data we code several important household events: birth of a child, household member retirement, household head job loss, departure of children ("empty nest"), divorce and marriage. In addition, we code changes in household income, as measured by changing income categories in the Nielsen panelist demographic survey. Panel B of Table 2 shows summary statistics on changes in household circumstances. Given the nature of the data collection we expect this to be measured with error, but we believe that these do provide some information on changes in household circumstances.

3 Event-Associated Changes in Diet

We begin by extending the literature on dietary changes on average, analyzing the impact of household events on diet in this sample.

3.1 Empirical Strategy

The events we consider occur at the household-year level. Our empirical strategy therefore 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, $\mathbf{T_{it}}$ is the vector of the treatment variable (years from event – diagnosis, household change), 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. The data includes households who do not experience an event, so even in the case of the disease diagnosis - which all occur in 2009 - we can separately identify year and treatment effects.

3.2 Results

Main Results Figure 1 shows the movement in diet score around disease diagnosis - hypertension in Figure 1a, obesity in 1b and diabetes in 1c. In the case of diabetes, the diet score improves slightly around diagnosis. The effect is significant, although small; about 0.1 of a standard deviation. As a benchmark, this is about a third of the cross-sectional difference in diet score between college educated households and those without a high school degree. For obesity and hypertension the changes are very small and not significant.

Panel A of Table 3 shows these effects statistically; because we are running many tests in this table, we indicate significance both with standard p-value cut-offs and with a Bonferroni correction. The results echo Figure 1. There are some small effects after a diabetes diagnosis, which are significant even when adjusting for multiple hypothesis testing. There is no consistent change in behavior after diagnosis with hypertension or obesity.

In Panel B of Table 3 we show evidence on household changes. After addressing issues of multiple hypothesis testing, the only significant changes are a short-run worsening of diet after childbirth and small improvements in diet after retirement and job loss. In the latter cases, however, it appears that pre-trends may be driving the results. These results are also extremely small. Divorce, marriage, departure of a child and even household income

increases do not seem to affect diet quality.

These results show behavior changes for an entire household; household-level analysis could mask larger individual changes. One way to consider this is to note that the average household size is 2.6. If all changes acrue to one individual, the size of the change could therefore be up to 2.6 times larger than the estimates. For diabetes, this would mark a somewhat sizable change. For the other changes, which are very small to begin with, it isn't clear that scaling up would make a large difference.

This multiplication approach likely over-estimates, though, since changes are likely to occur at the household level. As a further test, Appendix Table A3 shows the primary results for the subset of single-person households. We do not see sizable changes in this group; in some cases the effects are slightly smaller than we see in the overall sample, and in some cases they are slightly larger. Overall, this does not suggest that the null results are the effect of analyzing change at the household level.

Heterogeneity We can further expand this analysis to consider heterogeneity across demographic groups. In particular for something like disease diagnosis, a simple human capital theory would predict greater responsiveness among richer and more educated households, and among younger people. To explore this, we estimate - for each event - the responsiveness among household in the top tercile of education or income, and in the youngest tercile. Theory would predict these groups should respond more than the average person.[4]

The results are shown in Table 4, with each row representing an event. We aggregate the "before" and "after" periods to provide a summary measure of the impact. There is no strong evidence of heterogeneity in response. More educated, richer and younger households do not respond more to any of the events considered, at least not in a consistent way. Although we do see some limited differences (for example, higher income people respond more to income changes), we see equally compelling evidence in the other direction (lower income households respond more to diabetes diagnosis).

[4]We may also wonder about heterogeneity along the dimension of baseline diet quality - do those with worse diets change more? There are mechanical reasons to think this might be true and, indeed, it is true to some extent. However, the differences are not large.

4 Identifying Successful Diet Changes

The preceding section echoes the existing literature, showing diets appear to be on average unresponsive to changes in illness or household circumstances. We know from other literature that there are at least some individuals who do successfully improve their diet. In this section, we describe identification of households with dietary improvements in these data.

We define households with a successful diet improvement as those who change their diet score by an amount that is both large enough to represent a meaningful change in diet quality and is sustained. We call these households "changers".

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.[5] This leads to concerns about mean reversion.

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.

This problem of mean reversion is closely related to the discussion in Chay et al. (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 by incorporating two restrictions. First, we limit our data to a subsample 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

[5]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 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.

mean. Effectively, we try to begin with a sub-sample that is as close as possible to seeing a full measure of grocery purchases. This sample contains 23,766 households, versus 158,792 in the total sample. The significant limitation in sample size is due to the fairly stringent restrictions; households who (for example) have one month over a two and a half year period in which they are inconsistent scanners will be excluded.

Second, we identify household dietary changes based on a long pre-period and a long post-period. We require households to have an improvement in their diet score over a twelve month period, following a twelve 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 $c = 0.05$.

We choose the cutoff value of $c = 0.05$ to trade off a reasonably sized change with identification of a sizable number of households. The average change for this group in the year after versus before the change is 0.28, which is approximately three times as large as the college-high school difference in the cross-section.

In addition to this, we also classify households as changer households if they fulfill the criteria above with a single intermediate month between the two periods. This allows for households who change more slowly, or over periods which span the months we define. It might also be of interest to look at even slower changes, but given our data it will be hard to separate this type of changes from overall trends. We will therefore focus on this more narrow outcome of sudden and large change.[6]

We identify 1,613 changer households, approximately 6.75% of the sample. The remainder of the (non-changer) households will comprise our comparison group. We assign them a

[6]It may also be interesting to consider households who make a change for a shorter period and then revert. It is challenging, however, to separate this phenomenon from the more basic issue of reversion to the mean, which is a more significant problem as we limit the change period to be shorter.

random "change" date for the purposes of graphical comparison.

Figure 2 shows the movements in the diet score for the identified changer households (and the non-changers) over a period of 40 months.[7] The twenty-four months in the middle of the graph are the period based on which the households are chosen. During this period there is a large improvement in the diet score.

The difference is smaller if we compare the pre-pre-periods (months -24 to -12) to the post-post-period (months 12 to 24) but there is still a large difference in the diet score before and after. This suggests we are identifying changes which persist - at least to some extent - over time. To give a sense of the magnitude of these changes, in order to achieve a diet score improvement of this magnitude, the average household would have to reduce soda consumption by about 80 12-ounce cans per month, or 8 frozen pizzas.[8]

Appendix Figure A1 shows changes in other dietary measures for these households. These figures demonstrate reductions in the foods in unhealthy categories (soda, sugar, sweets, candy) and an increase in foods in healthy categories (fruits, vegetables, whole grains). These appear both in shares and in levels. There is also an improvement in the nutrient ratio, which is a second composite measure of diet quality based on nutrient values. In addition, we observe an increase in total food spend and a decrease in calories, likely reflecting the higher cost per calorie of healthy food.

Relative to the non-changer households, the changer households consume slightly more calories per person month (an average of about 66,000 versus 56,000) and are in slightly smaller households (21% of changers are in single-person households, versus 13% of the non-changers). The latter result may reflect the fact that changes for a single person may be more likely to be identified in a smaller household. We will explore these predictors more in Section 5.

[7]Note that although the middle 24 months are fully balanced, since we require only 30 months to be in the sample, the overall graph is slightly unbalanced.

[8]For these calibrations we examined baseline spend levels all changer households. Holding constant all other food category spending, we decreased each household's unhealthy spend by $1 at a time, and re-calculated the household's diet score. We stopped when the household's diet score increased by 0.28 points. On average this implied a reduction in unhealthy food spending of $33/month. Given an average price of $0.41 for a can of soda and $4.50 for a frozen pizza in the Nielsen data, this is equivalent to the stated reductions in those food categories.

Falsification Checks One concern is that these changes reflect not a true improvement in diet quality but instead an adjustment in scanning behavior. If households start scanning more, or less, we may incorrectly perceive that as a change in diet quality. We can look at this directly in the data.

Appendix Figure A2 shows evidence on scanning behavior at the time of the identified diet improvement. The first two graphs show purchases of non-food groceries and health products, which we take as one non-diet measure of scanning behavior. Although these are both trending, they do not show a large jump (in either direction) at the time of the change.

In addition, we can look directly at changes in number of trips recorded. To do this, we calculate the absolute value of the year-to-year change in number of trips in the pre-period, in the year surrounding the diet improvement, and the post-period. This is the third figure in Appendix Figure A2. Although on average there is some change in trip frequency for all households, the households with the diet improvement show no excess change in the year in which the improvement took place.

As a comparison, and a window into the power of this test, we show that if there *was* a fundamental change, these figures would pick it up. To do this, we create a set of fake melded households by combining two households together at an arbitrary date. Not surprisingly, a large share of these households would be identified as "changer" households, and show similar dietary improvements, etc. However, Appendix Figure A3 show the same falsification tests for these "fake" changers. These figures show large changes in demographics and scanning behavior. We take this as evidence that these plots would pick up changes if the diet improvement were driven by changes in the composition or scanning behavior of the households.

Finally, we consider what we may learn by looking for households who make changes in the opposite direction - worsening their diet. Our claim here is that the households who improve their diet are *choosing* to improve; we expect many fewer households will *choose* to worsen the quality of their diet. If the changes we observe are driven by changes in scanning behavior, however, we may expect to see a similar number of households worsening their diet.

To implement this, we look for household who worsen their diet score in a similar mag-

nitude to those who improve it. Not surprisingly, we do identify some households with this pattern, but we see only half as many as the improvement households. Further, when we look at our central placebo figure - on the changes in number of trips over time - for this sample, we see more evidence of changes in scanning behavior (Appendix Figure A4). Together, we view this as encouraging. Even if this share of "diet worsen-ers" is reflective of scanning changes, our diet improvement sample must contain a large share of people whose changes are not a result of scanning changes. Moreover, the fact that in this sample we *do* see some evidence of changes in scanning behavior which is not present in the diet improvement sample is an encouraging point about the power of that test and the argument that the diet improvements are not a result of scanning changes.

5 Analyzing Dietary Change

Using the identified changer households, this section asks two questions. First, to what extent can these households be predicted? Second, what are the patterns of behavior change?

5.1 Predicting Dietary Changes

We begin with prediction. We look at the predictive power of three categories of variables. First, events - those outlined in Section 3, along with more global events like changes in government dietary recommendation or research study releases[9]. This explores whether the timing of change is predictable. Second, baseline demographics: are some demographic groups more likely to change than others? Third, baseline diet features: do some types of diets seem more amenable to change than others?

5.1.1 Random Forest Learning

To optimize prediction, we use a machine learning model, specifically a random forest. The random forest is in the class of tree-based machine-learning prediction algorithms.[10] They

[9]In particular, we include the timing of key research study releases on the Mediterranean diet in 2008, 2009 and 2013, and the change in government dietary advice from a food pyramid to a food plate in 2011.

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

have been used elsewhere in economics, although not frequently (e.g. Oster, *forthcoming*; 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 has been shown to improve prediction.

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 and baseline diet characteristics. The diet characteristics include expenditures by category. In addition, we include a variety of measures of diet structure, including the "concentration" of the diet in various metrics. This concentration is measured by the standard deviation of the shares across Nielsen food groups, Nielsen food modules and the USDA "Thrifty Food Plan" groups (a slightly larger grouping).

We grow a random forest using 600 trees. We implement the random forest in R using the randomForest and randomForestSRC packages.[11] Further details on implementation are in Appendix B.

[11] 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.

5.1.2 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 using a Receiver Operating Characteristic (ROC) curve. 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 indicates 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.

We demonstrate the predictive power of the random forest using various sets of features. First, Figure 3a shows the ROC curve when we predict the outcome using only events. The AUC is 0.53, and we can see the line is very close to the 45 degree line. Effectively, we cannot predict changes in diet based on household events. This is consistent with our lack of effects in Section 3.

Figure 3b shows the quality of the prediction using only demographic information. This is slightly more predictive than events, but the predictive value is still fairly low; the AUC is 0.59.

Figure 3c shows the same curve using only baseline diet features. The prediction is much better; the AUC is 0.73. A useful way to summarize is to note that if we targeted the people with predictions in the top 10%, approximately 40% of them would be successful changers, versus 6% of the overall population.

Figure 3d shows the curve using all features. The AUC here is the same as the baseline diet alone. This suggests that even when incorporating events and demographics the primary predictive value is coming from the baseline diet.

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. Table 5 lists the top features in the full-feature forest, ranked by their importance.

These features are in two categories: measures of the quality of the diet, and measures of diet concentration. Baseline diet and share of foods in low quality diet groups are in the first set. The various measures of concentration - the standard deviation of shares across groups, the size of the largest share, etc. - are representative of the second. It is important to note that the exact foods or concentration measures which appear in this list are somewhat arbitrary. For example, removing baking supplies does not change the overall prediction quality. What we take from this is that these two categories - diet quality and concentration - seem to be the important drivers.

This importance list, however, misses some of what we are most interested in. 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.

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.[12] 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 for the entire group together in Figure 4. Panel (a) shows the relationship with baseline diet score and Panel (b) with the top measure of diet concentration (Thrifty Food Plan Groups) in terms of importance.[13] For baseline diet we observe a non-linear relationship. Dietary

[12]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.

[13]We note that we could make extremely similar graphs with any of the diet quality measures and any of the concentration measures.

improvement is more common for households with both better and worse diets at baseline.

In the case of concentration, the relationship is strongly upward sloping. Households with more concentrated diets at baseline are more likely to change. Intuitively, a more concentrated diet is one in which there are a few foods which make up a large share of purchases. It is worth saying that diet quality and conentration are correlated - a higher concentration is associated with worse diet quality - but the random forest suggests that they matter independently.

Overall, we take from this that behavior change does have some predictable features, and note the role of concentration as a key predictor.

5.2 Patterns of Behavior Change

Following prediction, the second question we ask is what are the patterns of behavior change among these households. How do they improve their diet?

We begin by showing, in the first two columns of Table 6, the food groups with the largest spending share changes on average. The foods with the largest reductions are frozen pizza, cookies, soda and ice cream. Those with the largest increases are yogurt, nuts, fruit and lettuce. The changes in these groups are substantial; frozen pizza, cookies and soda decline by 30 to 40% from their baseline, and consumption of the healthy food at least doubles in all cases.

This result masks significant heterogeneity across households in the largest change areas. Columns (3)-(4) of Table 6 show the share of households with each food group as their maximum change group. The changes are dispersed. Thirteen percent of households have frozen pizza as their maximum group, 12% have yogurt, 10% nuts.

Motivated by the findings on concentration above, a natural question to ask is to what extent the changes in diet within a household are also concentrated in a small number of foods. We analyze this by looking at each food group's contribution to the overall change in the diet score. We ask the question: if the household had *not* made the change in this particular food group, how much of the change in the diet score would be eliminated? Given that there is substantial heterogeneity across households in which particular food groups show the largest change, we structure this analysis to look at the importance of the largest

change group, the largest two change groups, etc., where the particular food in these groups varies across households.

These results are reported in Table 7. The two rows focus on three different food groupings - the top row shows the result when we focus on Nielsen product groups (largest), the middle for the aggregation based on doctor ranking and the bottom for module groups (smaller). In each case, column (1) reports the number of groups, and column (2) reports the average number of groups with positive spending in each year. In the case of product modules, in particular, there are a lot of modules which are so small that they are only infrequently purchased; the average household buys food in 118 of 484 possible modules.

Columns (3) - (5) show the share of the change in the overall diet score accounted for by the largest change groups - the top group, the top 2 and the top 5. When we look at product groups, where the average household buys items in 40 groups, the group with the largest change accounts for 60% of the overall change in diet score. Virtually all of the change is accounted for by the top two groups. For modules, which are much smaller, the top change module accounts for 30% of the change, and the top two for half. Nearly all of the change (82%) is accounted for by the top 5 modules.

One implication of this is that the large change households will begin with more concentrated diets, and their dietary concentration will decline after the change, as they make large changes in their highest consumption groups. Indeed, this appears in the data; we illustrate changes in concentration over time in Appendix Figure A5.

In the random forest output we observed that households with better diets and worse diets were, on average, more likely to improve their diets. A natural question is, therefore, whether the concentration patterns are different for those with a high quality versus low quality baseline diet. Table 8 replicates Table 7 for the top and bottom terciles of baseline diet. These patterns appear for both groups, but the changes are even more concentrated for those with a low quality diet at baseline. For this group, the top Nielsen food group accounts for 80% of the change in the diet score.

When we look at what foods they change, the households with a better diet at baseline are more likely to have their largest changes in healthy foods - nuts, fruit, yogurt. Those with a worse diet at baseline are more likely to change in frozen foods and soda. This is

perhaps consistent with the latter group having more scope to alter their unhealthy food purchases.

One question is whether these changes simply reflect the baseline purchase levels of these goods, in particular in cases where the diet improvements are driven by declines in unhealthy food group spending. Does this simply reflect a proportional decline beginning with a very concentrated diet? In fact, the reductions are more than proportional. Among households where the largest change is in unhealthy food groups, 63% of the change in diet score is accounted for by the food group with the largest changes. This group is, however, only 17% of the baseline food spending on average.

Overall, this section paints a picture of these successful diet households making a small number of sizable changes in their diet, rather than making small changes in many areas. Below we briefly discuss the implications of some of the data limitations for these results, and then turn to understanding how this, along with the other facts above, may fit within a model of behavior change.

6 Implications of Data Limitations

Before moving on to discuss what theoretical framework may help organize these facts, we pause to discuss more systematically how the limitations inherent in these data may bias the results above. The two key limitations are (1) the fact that the data is at the household level and (2) that we do not observe purchases of food away from home. We discuss these in turn below.

6.1 Household Level Data

We observe data at the household level only. We addressed this concern in the initial discussion of changes by showing that our null findings were robust to limiting to single-person households.

In the analysis of large change households we cannot do this for sample size reasons. It is therefore useful to think through what bias might be produced as a result of this.

Likely the central issue is that we may be more likely to identify single person households

as changers, since a large change among one person in a large household will be muted. We do see some evidence suggestive of this in the data; 21% of the large-changer household are single-person, versus 13% of those who do not change.

This suggests that we are missing some significant changes among larger size households so our overall estimate of the share of the population that makes changes like this may be an under-estimate. In addition, we may worry about the external validity of our results on concentration, if we think the effect of concentration is different across diffenet household types. We have no strong reason to believe this is the case, but it is challenging to test.

6.2 Food Away from Home

The Homescan data does not contain out-of-home food consumption, which is an important part of diet for most people, and tends to be less healthy than food consumed at home. This is a fundamental limitation of the data, and not a fixable one, but it is worth thinking about how it might bias our results.

First, in considering the baseline changes in response to diagnosis and other life events, we will underestimate responsiveness if households disproportionately respond by increasing the healthiness of their food away from home, or change their mix of home versus away. That is, if the main response to a diabetes diagnosis is to eat better out of the house, we will not pick that up here. Other related work (Oster, 2018) shows that in the particular context of diabetes changes in consumption are similar for foods which are nearly always eaten at home (i.e. breakfast cereals) suggesting that this problem may be limited.

In turning to the changers, it seems possible - even likely - that we have missed some large changers if some people improve their diet primarily or exclusively by limiting their food away from home. In this case we would identify only a subset of changer households, those who improve their diet through changes to their at-home consumption. When we then run analyses on this sample, we may not be learning about the broader population. This would threaten the external validity of the results (as with the household issue above), although not their internal validity.

A more pernicious issue in this large-changer identification is if we thought improvements in the health quality of grocery purchases were usually offset by a worsening in out-of-home

food consumption. In that case, we would not be identifying true changers, but simply people who altered the away-versus-at-home balance of their diet. It is important to be clear that since our metric is not of the *amount* of food they eat but the overall health quality, it would need to be the case that these households worsened the overall quality of their away from home diet. This offset story seems less likely to us, especially since the offeset would need to happen consistently over a long period. That is, a time-inconsistent phenomenon where someone tries to eat well at home and then is hungry so offsets with a fast-food dinner will affect our results only if the person continues to make this mistake for years.

In general, we anticipate that this limitations will cause us to miss some households who change their behavior, and probably miss some additional changes even among households we do identify. Seeing food away from home would be of value, in addition, since the predictive patterns we see in change suggest people change in a small number of categories. One of these could be, for example, food away from home. However, we believe that we do capture one important component of behavior change, or one important group of changers, and we can learn from that group.

7 Models of Behavior Change

The discussion above highlights several facts about changes in diet.

First, substantial changes in diet are rare, and their timing is unpredictable. Second, diet concentration predicts subsequent behavior change. Third, when households do make significant changes, these changes are concentrated in a small number of food groups.

In this section, we turn to the question of how these facts might be accommodated in a model of behavior change. We argue, first, that a simple neoclassical model will struggle to fit these facts together. We then argue that a model with attention costs may provide a better fit to the data. This section focuses on a simple model - one in which we assume convenient functional forms, etc., with the intention of illustrating one approach to fitting these data.

We should acknowledge, of course, that this is not the only model which can organize these facts.

Setup

A household chooses consumption from a menu of N possible foods, indexed by $i \in 1, ..., N$. We simplify the setup by assuming these are all unhealthy foods, which households would like to limit. Effectively, we consider dietary improvements to be reductions in bad foods, rather than increases in good foods. This accords with most of the data we observe.

Households consume x_i units of food i and consumption utility is given by

$$U = \sum_i \phi_i ln(x_i)$$

which specifies that households have concave utility in the consumption of each good, and the factor ϕ_i scales how valuable each good is to households. Without loss of generality we assume that $\sum_i \phi_i = 1$. Order the ϕ_i such that $\phi_1 > \phi_2 > ... > \phi_N$.

Each food i also has a health cost h_i per unit, and households have a health budget H such that

$$H = \sum_i h_i x_i$$

In this setup, H reflects the household's target level of dietary health. Note that a higher H implies a less healthy diet, since all foods in the model are unhealthy. In the analysis below, we consider the case of a household (possibly) reducing their health budget to $H - Z$. This captures an improvement in diet. We consider both the utility loss from making this change under various models and the optimal pattern of change across food groups.

Neoclassical Model We consider first results under a standard neoclassical model. Using standard constrained utility maximization, we find that with health budget H, households consume $x_i = \frac{H\phi_i}{h_i}$ at baseline for all i.

Proposition 1 below summarizes the changes when moving from a health budget of H to $H - Z$. Note that all proofs appear in Appendix C.

Proposition 1 *The optimal change in consumption for good i in moving from a health budget of H to $H - Z$ is $\Delta x_i = -\frac{\phi_i Z}{h_i}$ $\forall i$. The total utility loss is $U_{loss} = ln\left(\frac{H-Z}{H}\right)$.*

In this model, the optimal pattern of reduction in consumption is to reduce across all

foods consumed, in a manner that is proportional to baseline spend shares. The total utility loss is a function of the size of the change in diet.

We can consider the fit of this model to the data. The first fact - that people are unlikely to change their diet - can be accommodated here by assuming that the benefit to dietary change is smaller than the loss. Since we have not specified any benefits, this is straightforward.

However, the patterns of change observed are not consistent with this model. First, the utility loss is a function only of the size of the desired change, not of anything about the baseline diet. The observation that higher baseline concentration increases the chance of change is not accommodated here. Second, the optimal pattern of behavior change in this model is to reduce proportionally on all foods. However, we showed above that reductions are concentrated in a few foods, and are not proportional.

These facts suggest we may need to introduce some non-standard assumptions to the model in order to fit these facts.

Model with Attention Costs

We consider now introducing a model with an attention cost. This model draws - at least intuitively - on the "sparse-max" model of Gabaix (2014), which posits that bounded rationality may result from attention constraints. Our model is considerably simpler than the full sparse max operator, but draws on many of the same ideas. In line with that model, this attention model has underpinnings in the behavioral economics literature on two-systems and limited attention models (i.e. Kahneman, 2003; Gennaioli and Shleifer, 2010).

We introduce attention costs in a simple way. At baseline, households are consuming x_i units of each food i. We assume that households pay an attention cost m for each food group they change. Attention costs are constant whether the change in the food group is large or small. We should note that although we think of these as the costs of paying attention, this is isomorphic with any model in which there is a fixed costs to changing each individual food. This could arise from, i.e. habit formation.

The proposition below summarizes the patterns of change in this model.

Proposition 2 *(1) If changing fewer than N food groups it will be optimal to change the*

groups with the highest ϕ_i first; (2) for sufficiently high m it will be optimal to change fewer than N food groups; (3) if $m > (\phi_1 + \phi_2)\, ln\left(\frac{H(\phi_1+\phi_2)-Z}{H(\phi_1+\phi_2)}\right) - \phi_1 ln\left(\frac{H\phi_1-Z}{H\phi_1}\right)$ it will be optimal to change only group 1.

This proposition (and the proof in the appendix) shows that under this model the patterns of behavior change may be concentrated. Proposition 3 summarizes the relationship between dietary concentration and utility loss from change.

Proposition 3 *A household with a more concentrated baseline diet (captured by a higher ϕ_1 or ϕ_2) will have a lower utility cost from changing their diet than a household with a less concentrated baseline diet.*

Based on this proposition we observe that the probability of changing behavior at all is larger for households with a more concentrated diet. We note, however, that the addition of the attention cost distorts the behavior change such that the overall cost of changing behavior is considerably larger than in the neoclassical model.

Together, these results suggest this model does a better job fitting the facts. By introducing the idea of attention cost - something which is suggested in the existing literature - we can match the concentration of dietary change, and the observation that households with more concentrated diets are more likely to improve their diet. Intuitively, these households are able to make larger changes at lower cost since they can make more progress with a single food group.

This model also provides a natural way to think about why behavior change is challenging. There may be limited utility loss from making incremental changes in many food groups, but this is cognitively expensive. Households may be unable to make these changes since they are limited in their ability to devote cognitive resources.

To the extent that this model fits the data, it points to possible policy avenues. If households are challenged by making a large number of changes at once, it may be better to focus on making a smaller number of changes in high-value categories. Government dietary advice tends to focus on recommendations like "eat a balanced diet" or "eat fewer calories", which are non-specific. Under this model, it may be optimal to focus on specific changes (for example, limit or eliminate sugar-sweetened beverages). The focus on a smaller number

of simple changes may connect to the idea of the value of meta-rules when making hard changes, including in diet (i.e. Amir and Ariely, 2007; Payne and Barnett, 2018).

8 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 is a small share of households who do show large improvements in diet over time. Using a machine learning model we 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. The patterns of dietary improvement suggest the changes in diet are also largely concentrated in a small number of food groups.

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 evidence in the final section suggests that the data may be better fit with a model which takes into account households limited ability to pay attention to changes. This model better fits the facts than a standard neoclassical model, and suggests a focus on policy which targets a more limited number of changes to one's diet.

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.

Amir, On and Dan Ariely, "Decisions by rules: The case of unwillingness to pay for beneficial delays," *Journal of Marketing Research*, 2007, *44* (1), 142–152.

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.

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.

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.

Gabaix, Xavier, "A sparsity-based model of bounded rationality," *The Quarterly Journal of Economics*, 2014, *129* (4), 1661–1710.

Gennaioli, Nicola and Andrei Shleifer, "What comes to mind," *The Quarterly journal of economics*, 2010, *125* (4), 1399–1433.

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.

Hut, Stefan, "Determinants of Dietary Choice in the US: Evidence from Consumer Migration," 2018.

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.

Kahneman, Daniel, "Maps of bounded rationality: Psychology for behavioral economics," *American economic review*, 2003, *93* (5), 1449–1475.

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*, 2018, *10* (4), 308–48.

Payne, Christopher and Rob Barnett, *The Economists' Diet: The Surprising Formula for Losing Weight and Keeping It Off*, Vol. 1, Touchstone, 2018.

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.

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.

Table 1: **Summary Statistics**

Panel A: Panelist Demographics			
	Mean	*Standard Deviation*	*Sample Size*
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			
Diet Score	-0.296	0.193	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 Top 5 Groups:			
Carbonated Beverages	4.88%	6.48%	158,787
Bread and Baked Goods	4.8%	4.1%	158,787
Snacks	4.76%	4.57%	158,787
Fresh Produce	4.60%	4.59%	158,787
Milk Products	4.25%	4.8%	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: **Household Event Information**

A. Disease Diagnosis	*Total # Households*	*Any Diagnosis*	*New Diagnosis, 2009*
Hypertension/High Cholesterol/Heart Disease	33,369	52.1%	5.5%
Obesity	33,369	25.8%	1.3%
Diabetes	33,369	14.2%	0.7%
B. Household Changes	*# Household-Years*	*% With Event*	
Child Birth	150,153	1.3%	
Retirement	150,153	4.7%	
Head Job Loss	150,153	10.1%	
Empty Nest	150,153	0.9%	
Divorce	150,153	1.3%	
Income Increase	150,153	18.4%	

Note: This first rows of 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. The second panel of rows show the share of household-years with events indicating changes in family structure or circumstances.

Table 3: **Response to Household Events**

	(1) Yr -3 b/se/sd	(2) Yr -2 b/se/sd	(3) Yr -1 Mean	(4) Treat Yr b/se/sd	(5) Yr +1 b/se/sd	(6) Yr +2 b/se/sd
A. Disease Diagnosis						
Hypertension, cholesterol, heart	-0.001372	-0.004221	-0.280739	-0.000167	0.008782*	0.004464
	(0.00403)	(0.00315)		(0.00302)	(0.00382)	(0.00412)
	[-0.005]	[-0.016]		[-0.001]	[0.033]	[0.017]
Obesity	-0.001461	-0.006352	-0.282865	0.005475	0.005503	0.000020
	(0.00935)	(0.00730)		(0.00600)	(0.00809)	(0.00934)
	[-0.006]	[-0.024]		[0.021]	[0.021]	[0.000]
Diabetes	0.003801	0.005042	-0.301056	0.028719**+	0.040130**+	0.024665*
	(0.01020)	(0.00703)		(0.00887)	(0.00990)	(0.00982)
	[0.014]	[0.019]		[0.108]	[0.151]	[0.093]
B. Household Changes						
Child Birth	-0.016759*	-0.006518	-0.295624	-0.016047**+	-0.000940	0.008416
	(0.00724)	(0.00467)		(0.00362)	(0.00530)	(0.00722)
	[-0.063]	[-0.025]		[-0.061]	[-0.004]	[0.032]
Retirement	-0.007020*	-0.001770	-0.251780	0.005348**+	0.010832**+	0.014837**+
	(0.00295)	(0.00191)		(0.00169)	(0.00263)	(0.00364)
	[-0.027]	[-0.007]		[0.020]	[0.041]	[0.056]
Head Job Loss	-0.004322*	-0.000529	-0.294788	0.001920	0.005849**+	0.007117**+
	(0.00210)	(0.00140)		(0.00115)	(0.00165)	(0.00222)
	[-0.016]	[-0.002]		[0.007]	[0.022]	[0.027]
Income Increase	0.000693	-0.001414	-0.292502	-0.001305	-0.001977	-0.002600
	(0.00189)	(0.00124)		(0.00095)	(0.00134)	(0.00179)
	[0.003]	[-0.005]		[-0.005]	[-0.007]	[-0.010]
Empty Nest	0.013396*	0.007741	-0.327138	0.001785	0.005451	0.002595
	(0.00576)	(0.00402)		(0.00323)	(0.00458)	(0.00600)
	[0.051]	[0.029]		[0.007]	[0.021]	[0.010]
Divorce	0.004318	0.002581	-0.308072	-0.008261*	-0.011903*	-0.014093
	(0.00612)	(0.00409)		(0.00357)	(0.00532)	(0.00725)
	[0.016]	[0.010]		[-0.031]	[-0.045]	[-0.053]
Marriage	-0.001566	0.001899	-0.289302	0.004514	0.007962*	0.010324*
	(0.00423)	(0.00284)		(0.00235)	(0.00325)	(0.00428)
	[-0.006]	[0.007]		[0.017]	[0.030]	[0.039]

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 and household composition or circumstance changes on diet score. The outcome is the diet score we create based on doctor evaluations of the health of food groups. Disease diagnosis is drawn from the Nielsen Ailment Panel. Information on household circumstance changes is drawn from the yearly Nielsen panelist surveys.

Table 4: **Heterogeneity in Response to Events**

	(1)	(2)	(3)	(4)
	Avg	Education T3	Income T3	Youngest
A. Disease Diagnosis				
Hypertension, cholesterol, heart	0.00578	0.00563	0.00962*	0.00448
	(0.00307)	(0.00559)	(0.00490)	(0.00487)
Obesity	0.00849	0.00982	0.0160	0.00729
	(0.00668)	(0.0109)	(0.00947)	(0.0107)
Diabetes	0.0302**	0.0398*	0.0322**	0.0242*
	(0.00773)	(0.0175)	(0.0113)	(0.0116)
B. Household Changes				
Child Birth	-0.0151*	-0.0159	-0.0163*	-0.0157*
	(0.00593)	(0.00825)	(0.00707)	(0.00742)
Retirement	0.00604*	0.00561	0.00638	0.00730
	(0.00288)	(0.00440)	(0.0245)	(0.00379)
Head Job Loss	0.00401	0.00864**	0.00542	0.00886**
	(0.00206)	(0.00299)	(0.00325)	(0.00272)
Income Increase	0.0000367	0.00591*	0.00937**	0.00503*
	(0.00153)	(0.00236)	(0.00247)	(0.00211)
Empty Nest	0.00298	0.0155	0.00259	0.0132
	(0.00609)	(0.00845)	(0.00837)	(0.00802)
Divorce	-0.00365	0.00591	0.00443	-0.00292
	(0.00667)	(0.0101)	(0.00906)	(0.00913)
Marriage	0.0109*	0.0195**	0.0133*	0.0108
	(0.00435)	(0.00664)	(0.00628)	(0.00599)
Standard errors in parentheses				
Education tertiles: high school, some college, college/grad.				
P-values significance: * = 0.05, ** = 0.01				

Notes: This table shows the heterogeneity in the impact of disease diagnosis or household events by demographic groups. Column (1) shows the average response over the after period (year t to t+2) versus the before period (t-3 to t-1). Column (2) show the same for the highest education group (college or more), Column (3) for the top tercile of income and Column (4) for the youngest tercile of household head age group.

Table 5: **Important Features in Random Forest**

Baseline Diet Score
Share of purchases in baking supplies
SD of Shares, TFP groups
SD of Shares, Nielsen Modules
Frozen Entree Share
Baking mixes share
SD of Shares, Nielsen Groups
Largest Module Share
Fats and Condiments Share

Notes: This table shows the top importance features for the random forest. Standard deviations are measured in terms of shares across food groups. Groups are defined either by thrifty food plan (TFP) groups, Nielsen food groups or Nielsen Module Groups (smaller grouping).

Table 6: **Largest Diet Change Food Groups**

	Spending Changes, Top Items		***Top Changes, by Household Shares***	
Rank	**Name**	**PP Change Spending Share**	**Name**	**Top Item, % of HH**
1	Frozen Pizza	-3.4%	Frozen Pizza	13.5%
2	Yogurt	2.7%	Yogurt	12.0%
3	Nuts	2.4%	Nuts	10.1%
4	Fruit	2.4%	Fruit	9.2%
5	Cookies	-2.0%	Soda	7.4%
6	Soda	-2.0%	Cookies	5.9%
7	Ice Cream	-1.8%	Breakfast Bars	5.8%
8	Lettuce	1.5%	Ice Cream	5.2%

Notes: This table shows the important foods driving diet score changes. The first two columns show the groups with the largest changes on average. Column 2 reports the percentage point change in spending share for these groups. The second set of columns shows the distribution of the largest change group across households. For each item, we report the share of households who have that as their largest change group.

Table 7: **Concentration of Dietary Changes**

	Total	With Non-Zero Spend	Top 1	Top 2	Top 5
Nielsen Group	54	40	.601	.9433	1.407
Doctor Rank Groups	66	45	.5609	.889	1.42
Nielsen Module	484	110	.3221	.5027	.8195

Notes: This table shows the concntration of changes among the households who show large changes in diet. For each set of food categories, we report the total number and the household average number with non-zero spend. The last three columns report the share of the total change in diet score which is accounted for by the element of each category with the largest change, the top 2 and the top 5. Note that these figures can be greater than one if there are offsetting changes in other groups.

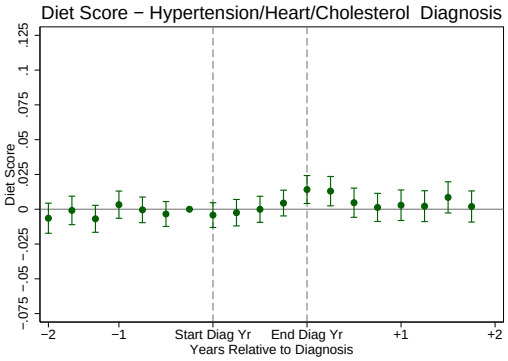
Table 8: **Concentration of Dietary Changes by Baseline Diet Quality**

Panel A: Top Tercile Baseline Diet					
	Total	With Non-Zero Spend	Top 1	Top 2	Top 5
Nielsen Group	54	41	0.52	0.79	1.20
Doctor Rank Groups	66	48	0.46	0.74	1.23
Nielsen Module	484	118	0.29	0.46	0.76
Panel B: Bottom Tercile Baseline Diet					
	Total	With Non-Zero Spend	Top 1	Top 2	Top 5
Nielsen Group	54	37	0.82	1.31	1.85
Doctor Rank Groups	66	40	0.69	1.06	1.65
Nielsen Module	484	93	0.39	0.60	0.94

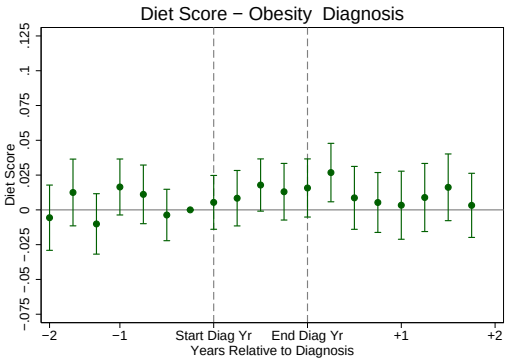
Notes: This table shows the concentration of changes among the households who show large changes in diet. Panel A includes households in the top third in terms of baseline diet quality. Panel B includes those in the bottom third. For each set of food categories, we report the total number and the household average number with non-zero spend. The last three columns report the share of the total change in diet score which is accounted for by the element of each category with the largest change, the top 2 and the top 5. Note that these figures can be greater than one if there are offsetting changes in other groups.

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

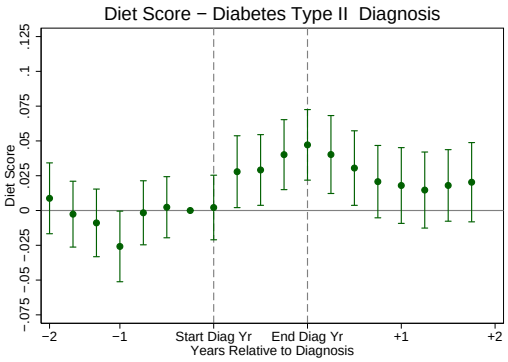
Panel A: Hypertension Diagnosis



Panel B: Obesity Diagnosis



Panel C: Diabetes Diagnosis



Notes: This figure shows the effect of diagnosis on the diet score of the purchased bundle for three diseases. 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: **Changes in Diet Score for Identified Changer Households**

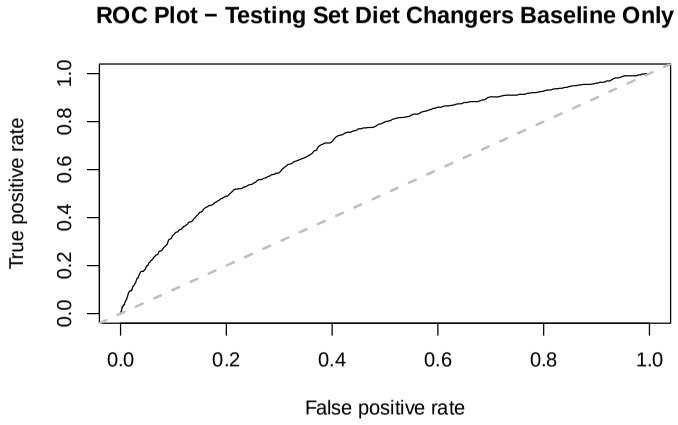
Notes: This figure shows the trend in the diet score for households identified as "large changers" and those who are not.

Figure 3: **Random Forest Output: Prediction Quality**

Panel A: Events Only

Panel B: Basline Demographics Only

Panel C: Baseline Diet Only

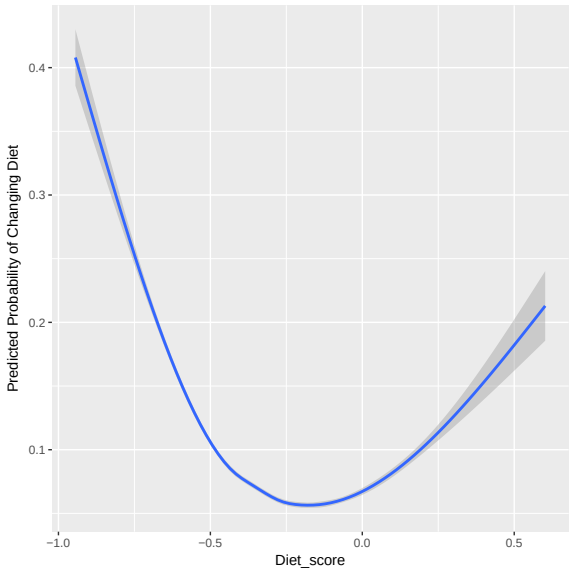


Panel D: All Features

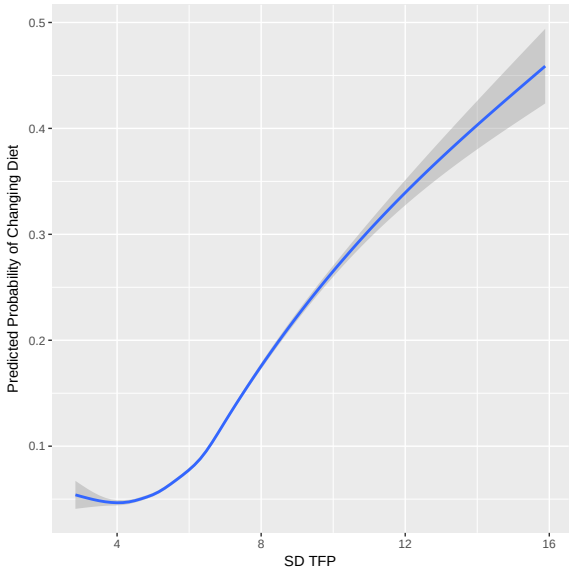
Notes: This figure shows the ROC curves from the random forest algorithm. Panel A uses only timing and events in the prediction. Panel B uses only baseline demographics. Panel C uses only baseline diet features. Panel D uses all features. The full list of features used in the random forest appears in Appendix C.

Figure 4: **Partial Dependence Plots for Random Forest**

Panel A: Role of Baseline Diet Score



Panel B: Role of Baseline Diet Concentration



Notes: This figure shows the partial dependence plots for the prediction of change in diet score. 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.

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

Notes: This table shows the demographics of the Nielsen sample relative to the US census in 2010. Standard errors in parentheses.

Table A2: **Comparison of Diet Score with Other Diet Measures**

	Diet score
Diet score	1
Unhealthy share	-0.462
Healthy share	0.482
Expenditure score	0.486
Nutrient ratio	0.273

Notes: This table shows the correlation between diet score and other measures of diet quality. Unhealthy share is the share of expenditures in soda, sugar, sweets and candies categories. Healthy share is the share of expenditures in vegetables, whole fruits and whole grains. Expenditure score is an overall measure of health of the diet (Handbury et al., 2015) and nutrient ratio is a measure based good and bad nutrient types in the diet (Drewnoski, 2005).

Table A3: **Response to Household Events, Single-Person Households**

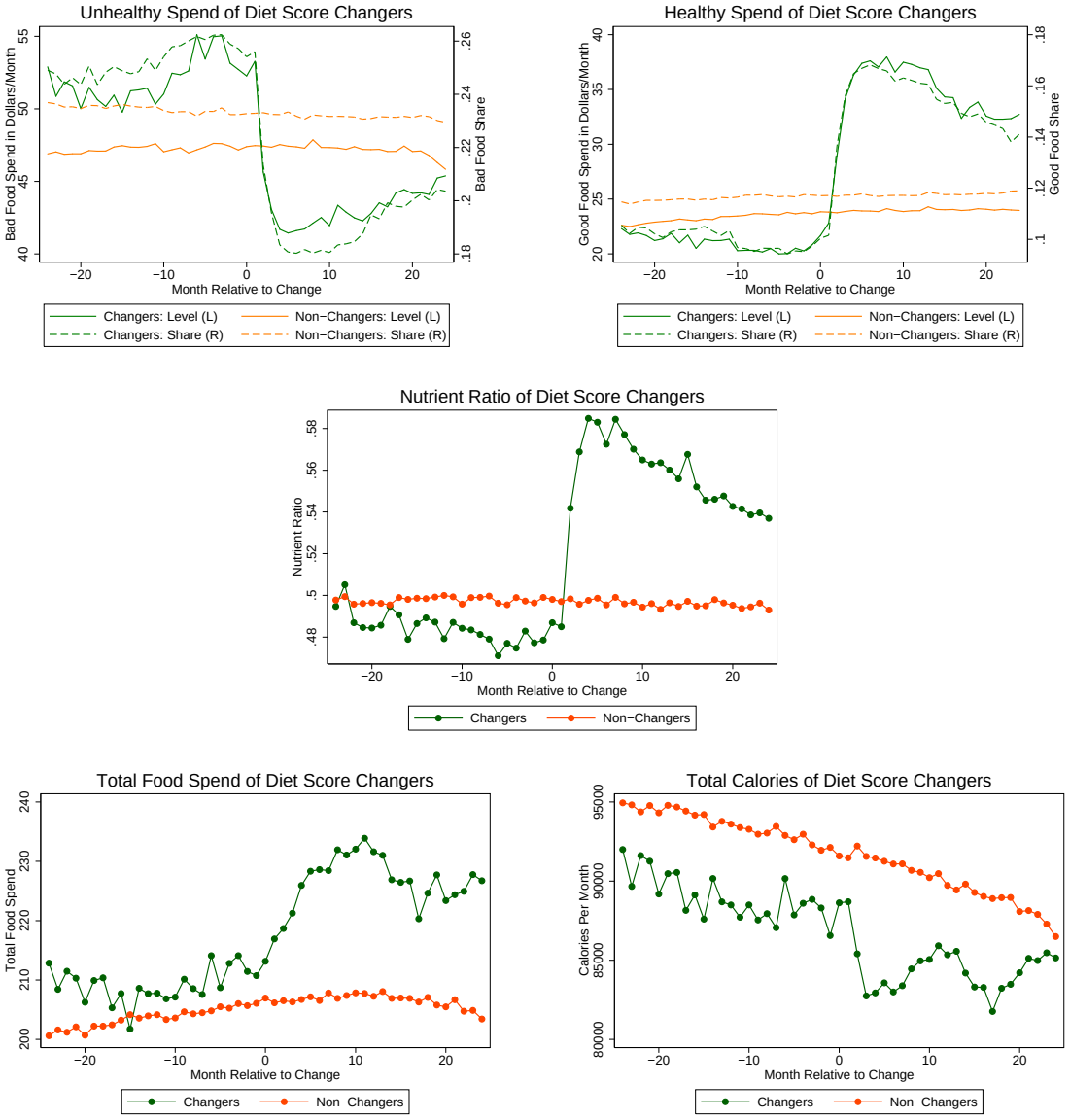
	(1) Yr -3 b/se/sd	(2) Yr -2 b/se/sd	(3) Yr -1 Mean	(4) Treat Yr b/se/sd	(5) Yr +1 b/se/sd	(6) Yr +2 b/se/sd
A. Disease Diagnosis						
Hypertension, cholesterol, heart	-0.002867	-0.008550	-0.288507	-0.007012	0.017670	0.010507
	(0.00906)	(0.00716)		(0.00738)	(0.00947)	(0.00948)
	[-0.009]	[-0.028]		[-0.023]	[0.057]	[0.034]
Obesity	-0.018295	-0.028223	-0.283973	-0.005891	0.005241	-0.008510
	(0.02139)	(0.01558)		(0.01333)	(0.01545)	(0.02117)
	[-0.059]	[-0.091]		[-0.019]	[0.017]	[-0.027]
Diabetes	0.005260	-0.015254	-0.283392	-0.003007	0.023673	0.019660
	(0.02626)	(0.01733)		(0.02056)	(0.02502)	(0.02475)
	[0.017]	[-0.049]		[-0.010]	[0.076]	[0.063]
B. Household Changes						
Retirement	-0.008957	-0.003716	-0.258295	0.006726	0.023102**+	0.018292**+
	(0.00618)	(0.00417)		(0.00367)	(0.00734)	(0.00556)
	[-0.029]	[-0.012]		[0.022]	[0.074]	[0.059]
Head Job Loss	-0.003989	0.003992	-0.295597	0.002997	0.014217*	0.008960*
	(0.00574)	(0.00405)		(0.00333)	(0.00562)	(0.00438)
	[-0.013]	[0.013]		[0.010]	[0.046]	[0.029]
Income Increase	0.003134	0.000619	-0.286684	-0.001191	-0.003483	-0.000407
	(0.00459)	(0.00296)		(0.00231)	(0.00465)	(0.00339)
	[0.010]	[0.002]		[-0.004]	[-0.011]	[-0.001]

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.

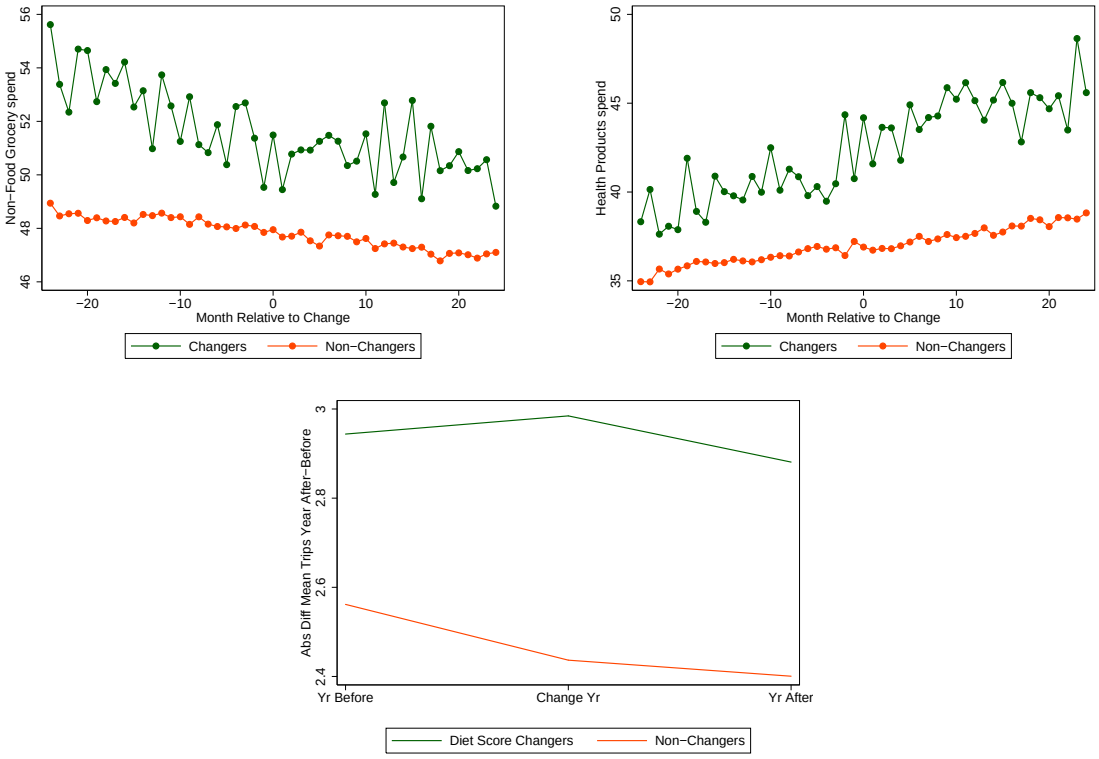
Notes: This table shows the impact of disease diagnosis and household composition or circumstance changes on diet score for single-person households. It can be compared to Table 3 in the main text, which shows the result for everyone. The outcome is the diet score we create based on doctor evaluations of the health of food groups. Disease diagnosis is drawn from the Nielsen Ailment Panel. Information on household circumstance changes is drawn from the yearly Nielsen panelist surveys.

Figure A1: **Auxiliary Changes in Large Changers:**



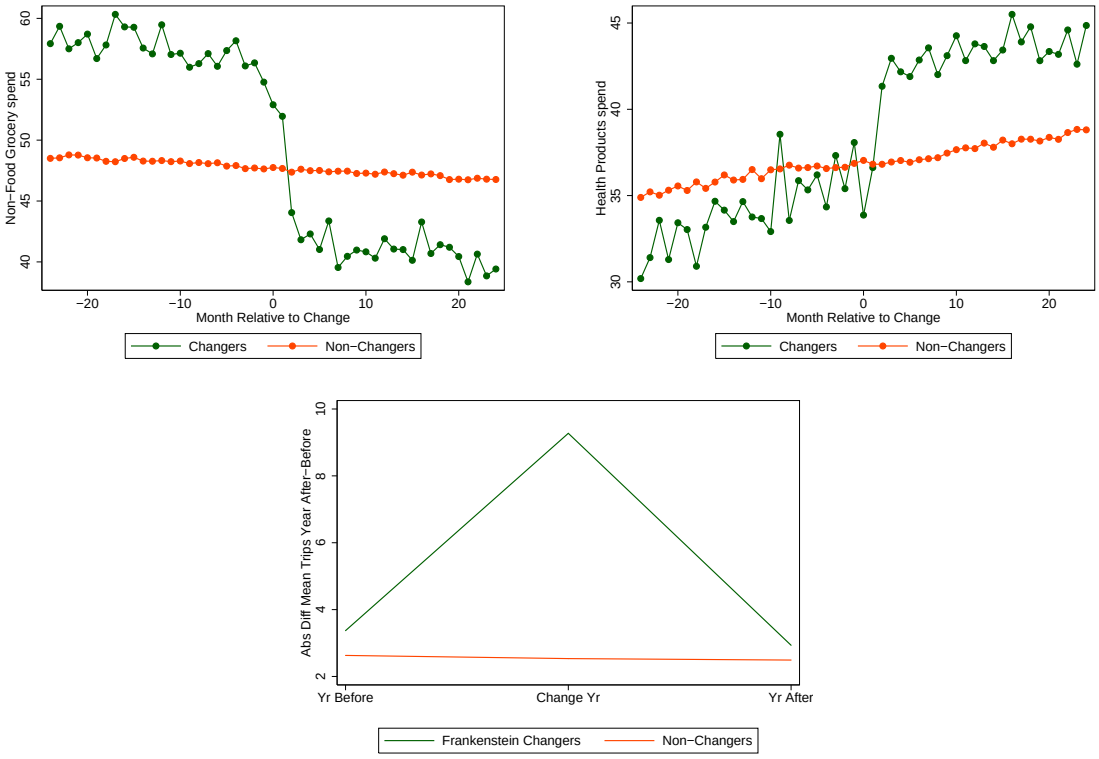
Notes: This figure shows the share of purchases in unhealthy foods and healthy foods, as well as changes in spending and calories for the households identified as diet score changers. Changer status is defined as a sustained improvement in diet score over a one year period following a one year baseline.

Figure A2: **Scanning Behavior, Large Changers**



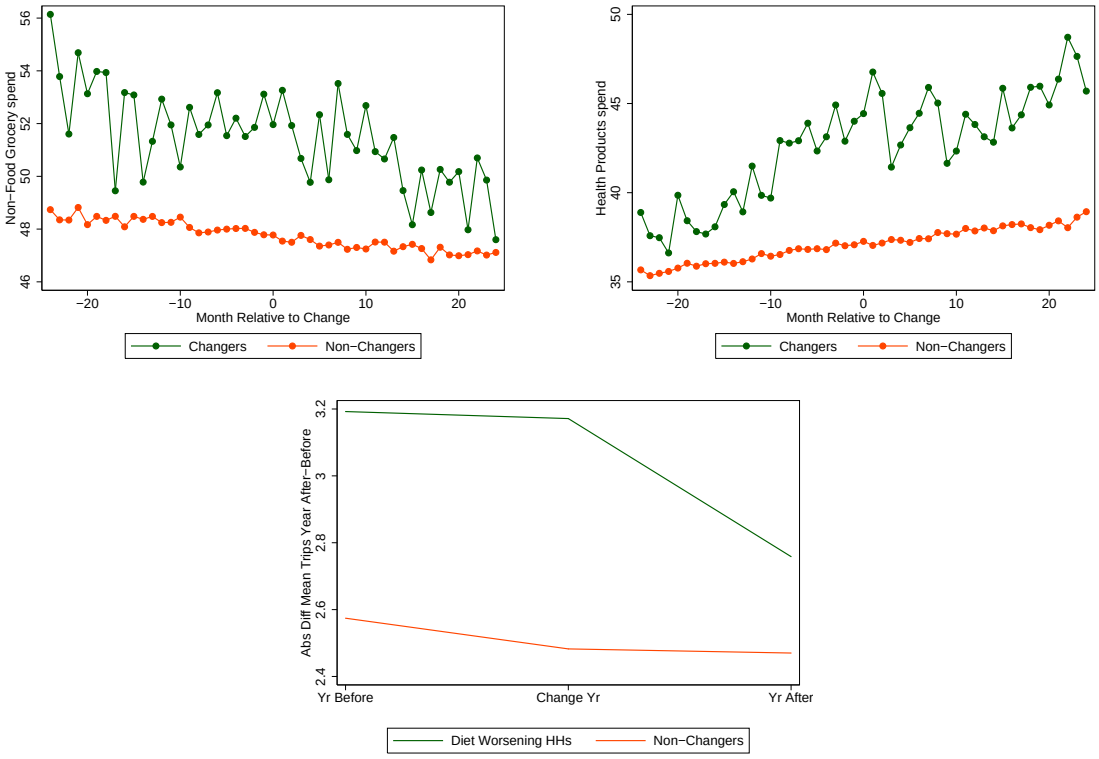
Notes: This figure shows several measures of scanning behavior for the households identified as diet score changers. Changer status is defined as a sustained improvement in diet score over a one year period following a one year baseline. The first two figures show purchases of non-food groceries and health products. The third figure shows the year-to-year change in average number of trips around the change date.

Figure A3: **Scanning Behavior in Constructed Household Large Changers**



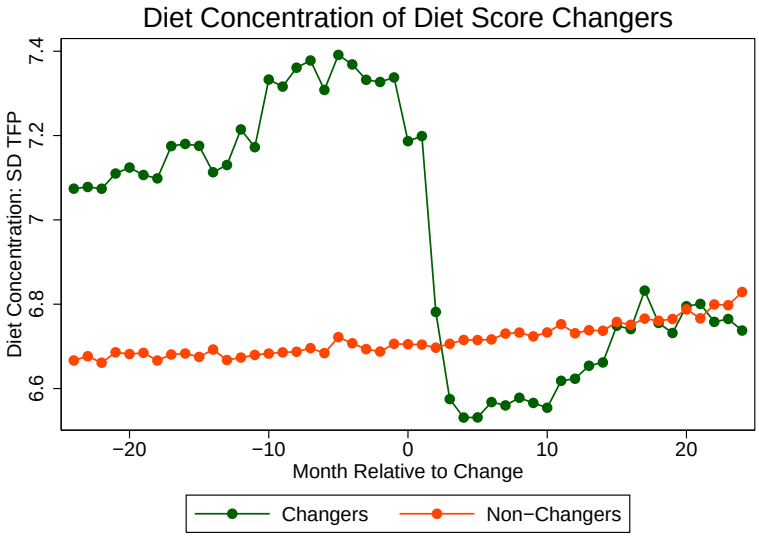
Notes: This figure shows scanning behavior changes among fake, composite households which are identified as large changers. We construct these households by appending two households together. We identify the households which show large improvements in diet quality as "changers". These are not true changers but cases where the household switches in the data. These figures show the corresponding changes in our measures of scanning behavior.

Figure A4: **Diet Worsening Households, Scanning Behavior**



Notes: This figure shows changes in scanning behavior among household who *worsen* their diet sharply. This accounts for approximately 800 households

Figure A5: **Changes in Dietary Concentration**



Notes: This figure shows changes in dietary concentration among our changer (and non-changer) households.

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 variable into 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.

Appendix C: Proofs

Proof of Proposition 1

We are looking for the solution to the following problem:

$$max_{x_i} U = \sum_i \phi_i ln(x_i)$$
$$s.t.$$
$$H = \sum_i h_i x_i$$

Note that we assume $\sum_i \phi_i = 1$. We can solve using a standard constrained maximization approach, and this yields the result that $x_i = \frac{H\phi_i}{h_i}$.

Proof of Proposition 2

This proof proceeds in 3 parts. First, we show that if you change only a subset of foods it is optimal to change the ones with the highest ϕ_i (Claim 1). Second, we show that it may be optimal to reduce fewer than N foods if m is sufficiently large (Claim 2). Third, we show that if m is sufficiently high it is optimal to change only one group (Claim 3).

Claim 1 We first show that if a household changes only one group it is optimal to change group 1 (with the highest ϕ_i). In order to reduce the health budget by Z changing only good i the household must change $\Delta x_i = \frac{Z}{h_i}$. The utility loss from this change is $U^i_{loss} = \phi_i ln\left(\frac{H\phi_i - Z}{H\phi_i}\right) - m$. The claim is that this loss is decreasing in ϕ_i. Since the loss is negative, this implies that the derivative of the loss function with respect to ϕ_1 is positive:

$$ln\left(\frac{H\phi_i - Z}{H\phi_i}\right) + \frac{Z}{H\phi_i - Z} > 0$$

The first part of this is negative, the second is positive. We know that $Z \in [0, H\phi_i]$ (so it is bounded between no reduction and completely eliminating x_i from the diet). On one side, as Z goes to zero, both pieces of this also go to zero, but the $ln(.)$ function converges more quickly to zero so the positive element dominates. As Z goes to $H\phi_i$ the $ln(.)$ goes to negative infinity and the second part goes to positive infinity. But since the $ln(.)$ function converges more slowly, the figure remains positive. Since this is positive on both extremes and is monotonic, it will be positive everywhere.

This shows that if only one group would be changed it would be the group with the highest ϕ_i.

Now consider a case where the attention cost is such that it makes sense to change more than one group, with one group changing Z_1 and the other Z_2. Note that the allocation of Z across the two groups will be a function of h_i and ϕ_i but it is clear this will be split. Assume $Z_1 > Z_2$. The optimal choice to reduce Z_1 is to reduce on the group with the highest ϕ_i (namely, group 1). To then further reduce Z_2 it is optimal to reduce on the group with the remaining highest ϕ_i (namely, Group 2). The logic continues through other groups. This completes the proof of this claim.

Claim 2 The utility loss from reducing all N foods is

$$U_{loss}^{all} = ln\left(\frac{H-Z}{H}\right) - Nm$$

The utilty loss from (optimally) reducing a subset j of foods is

$$U_{loss}^{subset} = \sum_{i=1}^{j} \phi_i ln\left(\frac{x_i - \Delta x_i}{x_i}\right) - jm$$

where $\sum_{i=1}^{j} \Delta x_i h_i = Z$.

It is trivial to observe that there are some values of m for which the subset loss would be lower, since $j < N$ and m is not bounded.

Claim 3 Claim (1) says that if only one group is changed it will be group 1. We show now that there is a value of m for which it is optimal to change Group 1 but not Group 2.

Note from above that the optimal baseline levels of x_1 and x_2 are $\frac{H\phi_1}{h_1}$ and $\frac{H\phi_2}{h_2}$, respectively. The loss from changing only good 1 is:

$$U_{loss}^{1} = \phi_1 ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right) - m$$

Denote the change in group 1 Δx_1 and note that $\Delta x_2 = \frac{Z - \Delta x_1 h_1}{h_2}$. We can solve for the optimal Δx_1 to minimize utility loss:

$$U_{loss}^{2} = \phi_1 ln\left(\frac{H\phi_1 - \Delta x_1 h_1}{H\phi_1}\right) + \phi_2 ln\left(\frac{H\phi_2 - Z + \Delta x_1 h_1}{H\phi_2}\right) - 2m$$

This yields

$$\Delta x_1 = \frac{Z\phi_1}{h_1(\phi_1 + \phi_2)}$$

And hence:

$$\Delta x_2 = \frac{Z\phi_2}{h_2(\phi_1+\phi_2)}$$

We can now compare the losses in the two cases. The loss from changing 1 will be smaller than two if :

$$\phi_1 ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right) - m >$$

$$\phi_1\left(ln\left(\frac{\phi_1\left(H(\phi_1+\phi_2)-Z\right)}{h_1(\phi_1+\phi_2)}\right) - ln\left(\frac{H\phi_1}{h_1}\right)\right) + \phi_2\left(ln\left(\frac{\phi_2\left(H(\phi_1+\phi_2)-Z\right)}{h_2(\phi_1+\phi_2)}\right) - ln\left(\frac{H\phi_2}{h_2}\right)\right) - 2m$$

This occurs if the following condition holds

$$m > (\phi_1+\phi_2)\, ln\left(\frac{H(\phi_1+\phi_2)-Z}{H(\phi_1+\phi_2)}\right) - \phi_1 ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right)$$

If this is true, then the only change made is in group 1.

Proof of Proposition 3

We start by showing this for the case when the household changes only one food group. The utility loss from reducing the health budget by Z when changing only one group is:

$$U^1_{loss} = \phi_1 ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right) - m$$

A diet is more concentrated the larger is ϕ_1. We proceed to show that the utility loss from changing one food group is increasing (less negative) in ϕ_1 :

$$\frac{\partial U_{loss}}{\partial \phi_1} = ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right) + \frac{H\phi_1^2}{H\phi_1 - Z}\cdot\left(\frac{Z}{H\phi_1^2}\right)$$

$$\frac{\partial U_{loss}}{\partial \phi_1} = ln\left(\frac{H\phi_1 - Z}{H\phi_1}\right) + \frac{Z}{H\phi_1 - Z} \geq 0$$

This is greater than or equal to zero as long as $H\phi_1 - Z \geq 0$, which has to be true because $Z \in [0, H\phi_1]$ (it is bounded between no reduction and completely eliminating x_1 from the diet).

For the case when a household changes two food groups, we use the optimal change levels

Δx_1 and Δx_2 derived in Claim 3 to find that the utility loss is:

$$Uloss = \phi_1 \left(ln \left(\frac{\phi_1 \left(H(\phi_1 + \phi_2) - Z \right)}{h_1(\phi_1 + \phi_2)} \right) - ln \left(\frac{H\phi_1}{h_1} \right) \right)$$

$$+\phi_2 \left(ln \left(\frac{\phi_2 \left(H(\phi_1 + \phi_2) - Z \right)}{h_2(\phi_1 + \phi_2)} \right) - ln \left(\frac{H\phi_2}{h_2} \right) \right) - 2m$$

A diet is more concentrated if$\phi_1$is higher, so we take the derivative with respect to ϕ_1 :

$$ln \left(\frac{H(\phi_1 H\phi_2) - Z}{H(\phi_1 + \phi_2)} \right) + \frac{Z}{H(\phi_1 + \phi_2) - Z} \geq 0$$

This is greater than or equal to zero as long as $H(\phi_1 + \phi_2) - Z \geq 0$, which has to be true because when changing two food groups $Z \in [0, H(\phi_1 + \phi_2)]$ (it is bounded between no reduction and completely eliminating x_1 and x_2 from the diet). Similarly, for a given ϕ_1, a diet is more concentrated if ϕ_2 is higher. It is easy to show that the derivative of the utility loss with respect to ϕ_2 is also positive.

The above shows that as diet is more concentrated at baseline, which occurs when $\phi_1$or ϕ_2 are higher, utility losses are smaller. Lastly, we demonstrate that the cutoff for changing only one food group is decreasing in ϕ_1.

$$\frac{\partial m}{\partial \phi_1} = \quad ln \left(\frac{H(\phi_1 + \phi_2) - Z}{H(\phi_1 + \phi_2)} \right) - ln \left(\frac{H\phi_1 - Z}{H\phi_1} \right) + \frac{Z}{H(\phi_1 + \phi_2) - Z} - \frac{Z}{H\phi_1 - Z}$$

We want to show that this derivative is weakly negative:

$$ln \left(\frac{H(\phi_1 + \phi_2) - Z}{H(\phi_1 + \phi_2)} \right) - ln \left(\frac{H\phi_1 - Z}{H\phi_1} \right) \leq \frac{Z}{H\phi_1 - Z} - \frac{Z}{H(\phi_1 + \phi_2) - Z}$$

We know that $Z \in [0, H\phi_1]$ since if a household changes one food group it can at most eliminate that food entirely from the diet. First, consider the case when $Z = 0$. In this case, the derivative is exactly equal to zero. Then consider the case where $Z = H\phi_1 - \epsilon$ for $\epsilon \to 0$. In this case the derivative is negative:

$$\underbrace{ln \left(\frac{H\phi_2 + \epsilon}{H(\phi_1 + \phi_2)} \right) - ln \left(\frac{H\phi_2 + \epsilon}{H(\phi_1 + \phi_2)} \right)}_{\to -\infty} \leq \underbrace{\frac{H\phi_1 - \epsilon}{\epsilon} - \frac{H\phi_1 - \epsilon}{H\phi_2 + \epsilon}}_{\to +\infty}$$

Lastly, we show that the derivative is decreasing in Z. Taking the derivative with respect to Z:

$$\frac{-1}{H(\phi_1+\phi_2)}+\frac{1}{H\phi_1-Z}-\frac{H\phi_1}{(H\phi_1-Z)^2}+\frac{H(\phi_1+\phi_2)}{(H(\phi_1+\phi_2)-Z)^2}$$

Simplifying:

$$\frac{Z}{(H(\phi_1+\phi_2)-Z)^2}-\frac{Z}{(H\phi_1-Z)^2}<0$$

This is negative as long as $\phi_2 > 0$.

Put together, these results demonstrate that both the number of food groups changed and the utility loss associated with that particular change are decreasing in baseline concentration. Therefore, a household with a more concentrated baseline diet faces an overall lower utility loss from changing behavior.