

Can nutritional label use influence body weight outcomes?

Andreas Drichoutis
Department of Economics
University of Ioannina, Greece
adrichou@cc.uoi.gr

Rodolfo M. Nayga, Jr.
Department of Agricultural Economics & Agribusiness
University of Arkansas, USA
rnayga@uark.edu

Panagiotis Lazaridis
Department of Agricultural Economics & Rural Development
Agricultural University of Athens, Greece
t.lazaridis@aua.gr



Can Nutritional Label Use Influence Body Weight Outcomes?

Andreas C. Drichoutis, Rodolfo M. Nayga Jr. and Panagiotis Lazaridis

Abstract

Nutritional labeling has been of much interest to policy makers and health advocates due to rising obesity trends. So can nutritional label use really help reduce body weight outcomes? This study evaluates the impact of nutritional label use on body weight using the propensity score matching technique. We conducted a series of tests related to variable choice of the propensity score specification, quality of matching indicators, robustness checks, and sensitivity to unobserved heterogeneity using Rosenbaum bounds to validate our propensity score exercise. Our results generally suggest that nutritional label use does not affect body mass index. Implications of our findings are discussed.

Keywords: Nutritional Labels, Body Mass Index, Propensity Score Matching, sensitivity analysis

JEL codes: I1, C14

I. INTRODUCTION

In light of the dramatic rise in obesity rates [particularly in developed countries (Loureiro and Nayga, 2005)], having healthy diets and healthier food choices are becoming the target of many public programs and policies. In the US, the Nutritional Labeling and Educational Act (NLEA) requires disclosure of the nutritional content of foods on a standardized label and strictly regulates the presence of health and nutrient content claims. The regulation also required a new format for the nutrition information panel and standardized serving sizes. Prior to the implementation of the NLEA, food manufacturers provided nutritional information only on a voluntary basis. More recently, due to the obesity issue, the provision of nutritional information in the food away from home (FAFH) market has also received a lot of attention in the US.

Nutritional labelling regulations are also being addressed in a number of countries around the world. For instance, the EU Commission consulted with member states and stakeholders in 2003 about the preparation of a proposal amending the voluntary provision of nutritional information to become mandatory. In November

2004, the Commission then published an impact assessment on the introduction of mandatory nutritional labelling for pre-packaged food products across the EU (European Advisory Services, 2004). As part of the consultation process, a paper discussing the revision of technical issues was also published in May 2006, paving the way for the final adoption of new mandatory rules.

Nutritional labeling regulations, both in the US and elsewhere, are aimed at helping people make more informed and healthier food choices. The literature for the US suggests that nutritional label use provides some dietary benefits. Specifically, increased use of nutritional labels has been associated with healthier patterns of dietary behaviour as well as food choice motivations (Coulson, 2000). Other studies associated label use with diets high in vitamin C and low in cholesterol (Guthrie et al., 1995) and with a lower percentage of calories from fat (Lin and Lee, 2003). In addition, disclosure of cereal brands' sugar content ("negative" information) caused consumers to switch to low-sugar cereals (Russo et al., 1986). Teisl *et al.* (2001) also found that food labelling can significantly affect consumer behavior. Although they did not find that providing health-related information always led consumers to switch consumption to 'healthy' products, others (Kim et al., 2001) have found that consumers' label use increased the average Healthy Eating Index (HEI)¹ by a range of between 3.5 and 6.1 points, with higher improvements in diet quality detected when health claim information was used. In addition, Variyam (2004) found that consumers who used the nutrition facts panel increased fiber, iron, and protein intakes, compared to consumers who did not use the nutrition facts panel. Neuhouser *et al.* (1999) found that label use was significantly associated with lower fat intake while Kim *et al.* (2000) found that label users generally had healthier diets than non-users, i.e., lower percentage of calories from fat and saturated fat, lower cholesterol and sodium intake, and higher fiber intake. Variyam (2008) also found that fiber and iron intakes of label users are higher than those of label nonusers.

As discussed above, a number of studies have evaluated the effect of nutritional label use on dietary outcomes. There is scant literature, however, on the effect of nutritional label use on body weight outcomes. Moreover, with a few exceptions [i.e.,

¹ The HEI is a measure of diet quality that assesses conformance to US federal dietary guidance. It was revised in 2006 to conform to the 2005 dietary guidelines for Americans and is now comprised of 12 components. HEI takes values from 0 to 100 where higher scores indicate a closer adherence to recommended ranges or amounts.

Kim *et al.* (2001) and Variyam (2008)], the important issue of self-selection problem inherent in the label use decision has been ignored in the literature. Label users and non-label users usually differ in observed socioeconomic and demographic characteristics. The bias caused when simply comparing the two groups and attributing the observed differences in health outcomes to label use alone is called self-selection because it ignores how individuals self-select into label use. Kim *et al.* (2001) addressed this problem by employing an endogenous switching regression model. Specifically, they compared the diet quality of label users and the expected diet quality of label users in the absence of labels. Variyam (2008) addressed self-selection by exploiting the fact that while nutrition information are mandatory for most foods sold in stores, food-away-from-home foods are exempt from the NLEA regulations. Therefore, label users - the treatment group - are exposed to the label in one setting but not in the other and label nonusers in the same settings act as the control group. Variyam then used a difference-in-differences estimator in this quasi-experiment.

A limitation in both studies, however, is the measure of label use they utilized in their analysis. In both studies, the ordinal measure of label use was collapsed into a binary indicator, thus setting *ad hoc* cut off points in the label use measurement. In this study, we relax this restriction by using an ordinal instead of a binary measure of label use. In addition, we explore the use of propensity score matching to address the possible occurrence of selection bias and reverse causality and to estimate treatment effects when treatment is endogenous to the outcome. Matching methods represent, depending on the particular method employed, either a semi-parametric or non-parametric alternative to linear regression (Black and Smith, 2004). The propensity score was introduced by Rosenbaum and Rubin (1983) to provide an alternative method for estimating treatment effects when treatment assignment is not random, but can be assumed to be unconfounded conditional on observables X . Since we argue that we observe the major variables influencing selection as well as outcomes, we assume that weight outcomes and selection into label use are independent conditional on these observables (Conditional Independence Assumption – CIA). We further examine and discuss the appropriateness of the CIA assumption later on in the paper.

Matching methods focus attention on a specific causal effect of interest and treat all variables other than the treatment variable as potentially confounding variables.

The influence of confounding variables is reduced by non-parametrically balancing the vector of characteristics across treatment, solely to obtain the best possible estimate of the causal effect of the treatment on the outcome variable. The idea is that people with the same propensity score that are on different experimental conditions (i.e. reading or not reading nutritional labels) can be compared after balancing the distributions of their covariates. Simply put, matching mimics a randomized experiment i.e. conditional on a set of observables there is some unspecified mechanism that randomizes people into treatment.

Since much of the debate on provision of nutrition information is founded on the obesity epidemic and since there are also some cries for policy interventions similar to the NLEA in the Food-Away-From-Home market, we examine the effect of label use on a weight outcome i.e. the Body Mass Index (BMI). In our propensity score matching exercise, we conduct a series of tests to validate our strategy. The tests relate to variable choice of the propensity score specification, quality of matching indicators and sensitivity analysis using “Rosenbaum bounds”. We also conducted robustness checks by estimating propensity scores for a “thick support” area (Black and Smith, 2004).

II. PROPENSITY SCORE MATCHING

Propensity score matching has become very popular in the estimation of causal treatment effects and has been applied to a wide variety of situations when there is a group of treated people and a group of untreated people. In this study, we depart from the binary treatment case since reading nutrition labels can occur at different frequencies which can be considered as different levels of treatment. Our aim is to assess the effect of each level of label use or treatment on BMI. However, we cannot observe all outcomes at the same time for the same individual and auxiliary methods are required. Taking just the difference of the mean outcomes between two levels of treatment would lead to selection bias since it is most likely that components which determine the treatment decision also determine the outcome variable of interest, and thus the outcomes of individuals from treatment and comparison group would differ even in the absence of treatment (Caliendo and Kopeinig, 2008). In what follows, we discuss propensity score matching for the binary treatment case and discuss later how we applied this to the multiple treatment case.

The idea behind the matching technique is to find a group of non-treated individuals that are similar to the treated individuals in all pre-treated characteristics X . That is, we construct an artificial comparison group and compare their health outcome (in terms of BMI) to label users. Under CIA, the matching estimator is consistent when the comparison group has the same distribution of observables determining health outcomes and selection as the label user group.

Most studies apply the propensity score matching technique in the case of binary treatments due to the wide availability of user written syntaxes that have made matching a simple estimation procedure. Generalizations in the case of multiple treatments (Imbens, 2000, Joffe and Rosenbaum, 1999, Lu et al., 2001) and continuous treatments (Hirano and Imbens, 2004) have also appeared but have not garnered much empirical attention yet. A workaround for the multiple treatment case was proposed by Lechner (2002), who employed several (matching) estimation methods for the multiple treatment case of active labor markets in the Swiss Canton of Zurich. He derived the probabilities used for the propensity scores from both a multinomial probit model and from all the possible binomial probits. He then compared the results and produced roughly the same answers.

Formally, in the binary treatment case, we assume that there is a variable T_i indicating treatment, which equals one if individual i uses nutritional labels (treated case) and zero if individual i does not use nutritional labels (control case). The propensity score is defined as the conditional probability of receiving a treatment (using nutritional labels) given pre-treatment (not using nutritional labels) characteristics X :

$$p(X) \equiv \Pr(T=1|X) = E(T|X) \quad (1)$$

If we define the health outcomes as H_{0i} and H_{1i} for the associated states 0 and 1, then the treatment effect for an individual i can be written as:

$$t_i = H_{1i} - H_{0i} \quad (2)$$

However, we do not know t_i for everyone since we can only observe

$$H_i = T_i H_{1i} + (1 - T_i) H_{0i} \quad (3)$$

i.e. either H_{0i} or H_{1i} . Since this problem cannot be solved at the individual level, it is recasted at the population level by estimating average treatment effects. The parameter,

which receives most attention in the literature is the average treatment effect on the treated:

$$t_{ATT} = E(t|T=1) = E(H_1|T=1) - E(H_0|T=1) \quad (4)$$

The problem with equation (4) is that the term $E(H_0|T=1)$ is not observed and if one tries to substitute this with $E(H_0|T=0)$, this would lead to “self-selection bias”.

The following assumptions are needed to derive (4) given (1) (see also (Caliendo and Kopeinig, 2005, Becker and Ichino, 2002, Heckman et al., 1998, Imbens, 2000):

Assumption 1. Balancing of pre-treatment variables

$$T \perp X \mid p(X) \quad (5)$$

Assumption 2. Unconfoundedness/ignorable treatment assignment (Rosenbaum and Rubin, 1983) or conditional independence (Lechner, 2002) or exogeneity (Imbens, 2004)

$$H_0, H_1 \perp T \mid X, \forall X \quad (6)$$

Assumption 3. Common support or overlap condition

$$0 < p(T=1|X) < 1 \quad (7)$$

Given assumptions 1, 2, and 3, the propensity score matching estimator is (Becker and Ichino, 2002, Caliendo and Kopeinig, 2005):

$$\begin{aligned} t_{ATT}^{PSM} &= E(E(H_1 - H_0|T=1, p(X))) \\ &= E(E(H_1|T=1, p(X)) - E(H_0|T=0, p(X)) \mid T=1) \end{aligned} \quad (8)$$

Equation (6) denotes the statistical independence of H_0 , H_1 and T on X and implies, that selection is solely based on observable characteristics and that all variables that influence treatment assignment and potential outcomes simultaneously are observed by the researcher (Caliendo and Kopeinig, 2005). Using the exact set of the observed variables as required for CIA to hold is a necessary step for the unbiased estimation of treatment effects. Rosenbaum and Rubin (1983), showed that when (6) and (7) are satisfied then $H_0, H_1 \perp T \mid p(X), \forall X$ which reduces the dimensionality of the matching problem substantially.

Assumption 3 has the unattractive feature that if the analyst has too much information about the decision of who takes treatment, so that $p(T=1|X) = 1$ or 0 the

method fails because people cannot be compared at a common X . The method of matching assumes that, given X , some unspecified randomization device allocates people to treatment (Heckman and Navarro-Lozano, 2004).

The method of matching with a known conditioning set does not require separability of outcome or choice equations, exclusion restrictions, or adoption of specific functional forms of outcome equations that are common in conventional selection methods and conventional instrumental variable formulations (Heckman and Navarro-Lozano, 2004). Furthermore, the method does not require exogeneity of conditioning variables. Lechner (2007) showed that it does not matter when some control variables may be influenced by the treatment as long as the usual formulation of the CIA holds. He then proposed an alternative formulation of the CIA together with explicit exogeneity conditions.

III. THE DATA

The data for our analysis come from the 2005-2006 National Health and Nutrition Examination Survey (NHANES), the latest available dataset. NHANES is designed to assess the health and nutritional status of adults and children in the US and is unique in that it combines interviews and physical examinations. The interview component includes demographic, socioeconomic, dietary, and health-related questions. The examination component consists of medical, dental, and physiological measurements, as well as laboratory tests administered by highly trained medical personnel.

Nutritional label use (i.e., use of Nutrition Fact Panels (NFP)) was measured on a five likert scale (never, rarely, sometimes, most of the time, always). Exploiting the full scale of the NFP variable requires recasting our propensity score matching exercise to the multiple treatment level case. However, even though there is an abundance of user written modules available to do matching for the binary treatment case, this is not the case for multiple treatments. A practical alternative as suggested by Lechner (2002) is to estimate a series of binomial models instead of modelling the joint selection process. The advantage is that a misspecification in one of the series will not compromise all others as would be the case in the multi-treatment model. The disadvantage is that the number of models to be estimated increases

disproportionately to the number of options i.e. for L options we need $0.5 L L-1$ models. Therefore, in our case for the 5 ordered treatments of the NFP reading variable, 10 binomial models need to be estimated. This means that each category is pairwise compared to all others (e.g. 5 vs. 4, 5 vs. 3, 4 vs. 2, 3 vs. 1 etc.). As indicated earlier, Lechner (2002) compared a multinomial probit with a series of binomial probits and found roughly the same answers/findings.

We utilize measured, not self-reported, body weight and height for our BMI measure. In general, the variables we use in our estimations are grouped into five categories: socio-demographic, risky behavior, lifestyle, knowledge and health situation variables. Although we realize that some of these control variables could possibly be endogenous, Lechner (2007) proved that this would not be a problem as long as the CIA holds. Socio-demographic variables include age, gender, race, education, household, size and income. Risky behavior variables consist of alcohol consumption, drug use, smoking status and engaging in safe sexual behaviour. Lifestyle variables consist of variables for Food-Away-From-home consumption, exercise frequency, perceived healthfulness of diet and food security of the household. Knowledge variables include variables that indicate if a doctor advised to reduce weight or eat less fat due to cholesterol problems or other chronic diseases, perceived knowledge of the Dietary Guidelines, the Food Guide Pyramid and the 5-a-Day program and a dietary variable indicating self-efficacy (“Some people are born to be fat and some thin; there is not much you can do to change this.”). Health Situation variables include pregnancy status, diabetes status, chronic diseases status and intake of diabetic medicine status. Observations with missing values for the variables of interest were dropped from all subsequent analysis. The sample size of our analysis is 4346. Descriptions of the variables used in our analysis are exhibited in Table 1.

Table 1. Variable description

	<i>Variables</i>	<i>Variable description</i>	Mean	Std. Error
	<i>BMI</i>	<i>Body Mass Index</i>	28.81	6.79
	<i>LabelUse₁[*]</i>	<i>Dummy, Never reads Nutrition Fact Panels</i>	0.32	0.47
	<i>LabelUse₂</i>	<i>Dummy, Rarely reads Nutrition Fact Panels</i>	0.099	0.30
	<i>LabelUse₃</i>	<i>Dummy, Sometimes reads Nutrition Fact Panels</i>	0.22	0.42
	<i>LabelUse₄</i>	<i>Dummy, Most of the time reads Nutrition Fact Panels</i>	0.19	0.39
	<i>LabelUse₅</i>	<i>Dummy, Always reads Nutrition Fact Panels</i>	0.17	0.37
Demographic	<i>Gender</i>	<i>Dummy, Gender of the respondent</i>	0.48	0.50
	<i>Age</i>	<i>Age of the respondent</i>	47.32	18.50
	<i>Race₁[*]</i>	<i>Dummy, Hispanic race</i>	0.22	0.42
	<i>Race₂</i>	<i>Dummy, Ethnicity is non-Hispanic White</i>	0.51	0.50
	<i>Race₃</i>	<i>Dummy, Ethnicity is non-Hispanic Black</i>	0.23	0.42
	<i>Race₄</i>	<i>Dummy, Other ethnicity</i>	0.04	0.20
	<i>Educ₁[*]</i>	<i>Dummy, High school grad/GED or equivalent</i>	0.50	0.50
	<i>Educ₂</i>	<i>Dummy, Some College or Associate of Arts degree</i>	0.29	0.45
	<i>Educ₃</i>	<i>Dummy, College graduate or above</i>	0.21	0.40
	<i>Hsize</i>	<i>Household size</i>	3.08	1.62
	<i>Inc₁[*]</i>	<i>Dummy, Annual household income < \$24,999</i>	0.30	0.46
	<i>Inc₂</i>	<i>Dummy, \$25,000 < Annual household income < \$54,999</i>	0.34	0.47
	<i>Inc₃</i>	<i>Dummy, Annual household income > \$55,000</i>	0.36	0.48
	Risky behavior	<i>DrinkDay</i>	<i>Average number of alcoholic drinks per day consumed over the past 12 months</i>	0.10
<i>DrugUser</i>		<i>Dummy, Respondent has used during the last month either of: hashish, marijuana, cocaine, heroin, methamphetamine</i>	0.08	0.27
<i>NoSmoke</i>		<i>Dummy, Respondent doesn't smoke</i>	0.25	0.43
<i>SafeSex</i>		<i>Dummy, Respondent has never had sexual intercourse without a condom, over the past 12 months</i>	0.10	0.30
Lifestyle	<i>MealsFAFH</i>	<i>Number of meals per week not prepared at home</i>	3.25	3.61
	<i>MET</i>	<i>Total Metabolic Equivalent rate of activities</i>	8.57	12.04
	<i>HealthyDiet₁[*]</i>	<i>Dummy, Respondent rates overall diet as poor</i>	0.06	0.24
	<i>HealthyDiet₂</i>	<i>Dummy, Respondent rates overall diet as fair</i>	0.23	0.42
	<i>HealthyDiet₃</i>	<i>Dummy, Respondent rates overall diet as good</i>	0.39	0.49
	<i>HealthyDiet₄</i>	<i>Dummy, Respondent rates overall diet as very good</i>	0.22	0.42
	<i>HealthyDiet₅</i>	<i>Dummy, Respondent rates overall diet as excellent</i>	0.09	0.29
	<i>FoodSecur₁[*]</i>	<i>Dummy, Household's food security is low or very low</i>	0.14	0.34

	<i>FoodSecur₂</i>	<i>Dummy, Household's food security is marginal</i>	0.09	0.28	
	<i>FoodSecur₃</i>	<i>Dummy, Household's food security is full</i>	0.77	0.42	
Knowledge	<i>DoctAdv₁</i>	<i>Dummy, Doctor instructed to eat less fat for cholesterol</i>	0.22	0.41	
	<i>DoctAdv₂</i>	<i>Dummy, Doctor instructed to reduce weight for cholesterol</i>	0.15	0.36	
	<i>DoctAdv₃</i>	<i>Dummy, Doctor instructed to eat less fat to lower the risk for certain diseases</i>	0.24	0.43	
	<i>DoctAdv₄</i>	<i>Dummy, Doctor instructed to reduce weight to lower the risk for certain diseases</i>	0.28	0.45	
	<i>KnowDG</i>	<i>Dummy, Respondent has heard of Dietary Guidelines</i>	0.43	0.50	
	<i>KnowFGP</i>	<i>Dummy, Respondent has heard of Food Guide Pyramid</i>	0.71	0.45	
	<i>Know5AD</i>	<i>Dummy, Respondent has heard of 5-a-Day program</i>	0.46	0.50	
	<i>Born2beFat₁*</i>	<i>Dummy, Respondent strongly disagrees that some people are born to be fat</i>	0.28	0.45	
	<i>Born2beFat₂</i>	<i>Dummy, Respondent somewhat disagrees that some people are born to be fat</i>	0.25	0.43	
	<i>Born2beFat₃</i>	<i>Dummy, Respondent neither agrees or disagrees that some people are born to be fat</i>	0.12	0.32	
	<i>Born2beFat₄</i>	<i>Dummy, Respondent somewhat agrees that some people are born to be fat</i>	0.24	0.43	
	<i>Born2beFat₅</i>	<i>Dummy, Respondent strongly agrees that some people are born to be fat</i>	0.11	0.32	
	Health Situation	<i>Pregnant</i>	<i>Dummy, Respondent is pregnant</i>	0.07	0.26
		<i>DocDiabet</i>	<i>Dummy, Respondent was told by a doctor that has diabetes, prediabetes or at risk for diabetes</i>	0.24	0.43
		<i>DiabMedicine</i>	<i>Dummy, Respondent takes either insulin or diabetic pills</i>	0.09	0.28
<i>Chronic</i>		<i>Dummy, Respondent suffers from coronary heart disease, heart attack, stroke or liver condition</i>	0.11	0.31	

These variables were removed for estimation purposes.

IV. ESTIMATION AND RESULTS

1. Plausibility of CIA and propensity score estimation

Before we proceed to the estimation of the propensity score, we have to support the plausibility of CIA for our case. For CIA to be fulfilled, one has to condition on all variables that simultaneously influence the participation decision and the outcome variable. Although CIA is a strong assumption, given that we have an extremely rich and informative dataset that allows us to control for a wide variety of socio-demographic variables, risky behavior, lifestyle, knowledge and current health situation, we argue that the CIA holds. Furthermore, in a latter section, we conduct sensitivity

analysis to explore how sensitive our estimates are to potential failures of the CIA assumption.

As far as the estimation of the propensity score is concerned, there is no clear cut rule on which variables to include in the treatment equation or on the functional form of the probabilistic model. Regarding the choice for the latter, usually a probit or a logit model is estimated. Lechner (2002) compared binary probit models with a multinomial probit and concluded that results between the models are roughly the same. Given the absence of comparisons between logistic and multinomial logit models, we then proceed by estimating binary probits.

Regarding the probit specifications, there are some formal statistical tests which can be used. Two such tests are the “hit-or-miss” method or prediction rate metric (Heckman et al., 1997) and the pseudo- R^2 . The latter indicates how well the regressors X explain the participation probability. With the hit-rate, variables are chosen to maximize the within-sample correct prediction rates, assuming that the costs for misclassification are equal for the two groups. The method classifies an observation as ‘1’ if the estimated propensity score is larger than the sample proportion of persons taking treatment, i.e. $\hat{P}(X) > \bar{P}$, and as ‘0’ otherwise.

Both of these statistics have been estimated for several specifications of all 10 binary models. As mentioned in the previous section, variables were grouped into one of the following five categories: socio-demographic, risky behaviour, lifestyle, knowledge and health situation. The base specification includes variables from only one of the aforementioned categories. Then all possible combinations of two, three and four categories and of the full specification are tested. In all, we tested 31 different specifications. Based on the pseudo- R^2 , the full specification does the best job in explaining the participation probability in all models. Given that in most cases the hit-rates from the full specification are equivalent to other specifications and since there is no economic justification in excluding categories of variables, we use the full specification².

² The reader should be aware that relying solely on goodness-of-fit criteria is not without warnings. Heckman and Navarro-Lozano (2004) offer examples where application of goodness-of-fit criteria point to selection of conditioning sets that are less successful in terms of a model selection criterion. However, these are still offered as possible solutions in the literature (Black and Smith, 2004; Caliendo and Kopeinig, 2008; Heckman, Ichimura and Todd, 1997) and the reader should not take such tests at face value.

Hence, the selection between treatment levels is expressed through the function:

$$T = f \left(\begin{array}{l} \textit{Gender}, \textit{Age}, \textit{Race}_2, \textit{Race}_3, \textit{Race}_4, \textit{Educ}_2, \textit{Educ}_3, \textit{Hsize}, \textit{Inc}_2, \textit{Inc}_3, \textit{DrinkDay}, \\ \textit{DrugUser}, \textit{NoSmoke}, \textit{SafeSex}, \textit{MealsFAFH}, \textit{MET}, \textit{HealthyDiet}_2, \textit{HealthyDiet}_3, \\ \textit{HealthyDiet}_4, \textit{HealthyDiet}_5, \textit{FoodSecur}_2, \textit{FoodSecur}_3, \textit{DoctAdv}_1, \textit{DoctAdv}_2, \\ \textit{DoctAdv}_3, \textit{DoctAdv}_4, \textit{KnowDG}, \textit{KnowFGP}, \textit{Know5AD}, \textit{Born2beFat}_2, \\ \textit{Born2beFat}_3, \textit{Born2beFat}_4, \textit{Born2beFat}_5, \textit{Pregnant}, \textit{DocDiabet}, \\ \textit{DiabMedicine}, \textit{Chronic} \end{array} \right) \quad (9)$$

Variables used in (9) are described in Table 1. Since we controlled for an exhaustive list of variables in equation (9), we expect to have very minimal unobserved heterogeneity left, if any, that is systematically correlated with the health outcome under investigation (BMI) and participation decision. In a later section, however, we tested the sensitivity of our results on possible unobserved heterogeneity or hidden bias.

The results from the propensity score estimations (we used the probability weights provided with the NHANES dataset) are summarized in Table 2. The “ k vs. n ” model (where $k=1$ to 5 and $n=1$ to 4) denotes the pairwise comparison of the k level of the NFP use to the n level of NFP use. For example, the “5 vs. 4” compares those that use the NFP label “*always*” ($k=5$) with those that use the NFP label “*most of the time*” ($n=4$). With respect to the socio-demographic variables, we find that males (*Gender*) and non-Hispanic white individuals (*Race*₂) are less likely to read NFPs. Education and income do have the expected effect i.e. the more education and the higher the income of individuals are, the more likely they are to read the NFP. Household size on the other hand is negatively related to NFP reading.

Some variables that were used to capture risky behaviour also have significant effects. For example, there is some indication that drug users are less likely to read NFPs while non-smokers are more likely to read the nutrition panel. With respect to lifestyle variables, we find that higher metabolic equivalent rates (*MET*), higher perceived healthiness of diet and higher household food security are all positively related to NFP reading probability.

Knowledge has a positive effect on NFP reading as well. Although not all variables are statistically significant, there is some indication that knowledge as expressed through doctors’ advice and perceived knowledge of the Dietary Guidelines, the Food Guide Pyramid and the 5-a-Day program are all related to NFP reading. In addition, agreeing to the dietary attitude that “some people are born to be fat and that

there is nothing you can do to change that” is negatively related to the probability of reading the NFP. Current health situation is also a good predictor of NFP search behaviour. A diabetic condition i.e. diagnosed with diabetes or taking insulin or pill to control it, positively affects NFP reading. Interestingly, pregnant women are less likely to read NFP’s.

Table 2. Probit models for label use

<i>Variables</i>	Models									
	5 vs. 4	5 vs. 3	5 vs. 2	5 vs. 1	4 vs. 3	4 vs. 2	4 vs. 1	3 vs. 2	3 vs. 1	2 vs. 1
<i>Gender</i>	0.014	-0.312**	-0.644**	-0.757**	-0.342**	-0.733**	-0.845**	-0.353**	-0.565**	-0.246**
<i>Age</i>	-0.004	0.002	0.005	0.002	0.004	0.008**	0.003	0.001	-0.001	0.001
<i>Race₂</i>	-0.524**	-0.265**	-0.292**	-0.223*	0.208*	0.116	0.187	-0.078	-0.066	-0.078
<i>Race₃</i>	-0.254*	-0.083	-0.195	-0.113	0.099	-0.092	0.032	-0.152	-0.121	-0.053
<i>Race₄</i>	-0.494**	-0.152	-0.290	-0.424**	0.304	0.156	-0.066	-0.098	-0.327*	-0.362*
<i>Educ₂</i>	0.054	0.148	0.187	0.384**	0.102	0.161	0.352**	0.066	0.291**	0.232**
<i>Educ₃</i>	-0.160	0.059	0.270*	0.552**	0.228**	0.479**	0.727**	0.196	0.494**	0.322**
<i>Hsize</i>	-0.042	-0.084**	-0.079**	-0.105**	-0.058**	-0.048	-0.099**	-0.015	-0.048*	-0.027
<i>Inc₂</i>	-0.009	0.109	-0.063	0.167*	0.100	-0.027	0.180*	-0.171	-0.007	0.149
<i>Inc₃</i>	-0.129	-0.058	-0.250*	0.284**	0.036	-0.178	0.324**	-0.183	0.297**	0.447**
<i>DrinkDay</i>	-0.109	0.034	0.130	0.014	0.061	0.246*	0.018	0.060	-0.066	-0.196
<i>DrugUser</i>	-0.013	-0.238	-0.260	-0.091	-0.261*	-0.328*	-0.139	-0.036	0.032	0.060
<i>NoSmoke</i>	0.064	0.203**	-0.021	0.193**	0.105	-0.092	0.144	-0.121	0.005	0.133
<i>SafeSex</i>	0.008	-0.075	0.017	-0.049	-0.102	0.108	0.058	0.158	0.013	-0.096
<i>MealsFAFH</i>	-0.009	-0.009	0.004	0.019	-0.003	0.004	0.024**	0.006	0.020**	0.022*
<i>MET</i>	-0.001	0.007**	0.012**	0.008**	0.009**	0.015**	0.008**	0.007*	0.004	0.000
<i>HealthyDiet₂</i>	-0.179	-0.131	-0.119	0.352*	0.038	0.059	0.460**	-0.027	0.468**	0.383**
<i>HealthyDiet₃</i>	-0.005	0.164	0.256	0.810**	0.192	0.297	0.846**	0.102	0.697**	0.568**
<i>HealthyDiet₄</i>	0.148	0.628**	0.672**	1.274**	0.514**	0.545**	1.129**	0.077	0.789**	0.538**
<i>HealthyDiet₅</i>	0.830**	1.115**	1.095**	1.343**	0.369	0.470	0.696**	0.141	0.405**	0.187
<i>FoodSecur₂</i>	0.074	0.183	0.227	-0.080	0.102	0.095	-0.156	-0.038	-0.197	-0.228
<i>FoodSecur₃</i>	0.159	0.381**	0.370*	0.163	0.174	0.078	0.045	-0.008	-0.061	-0.064
<i>DoctAdv₁</i>	0.041	0.047	0.345	0.209	0.041	0.430**	0.317**	0.352**	0.184	-0.207
<i>DoctAdv₂</i>	-0.061	-0.001	-0.157	0.027	0.103	-0.140	-0.011	-0.120	0.132	0.204
<i>DoctAdv₃</i>	0.000	-0.109	-0.129	0.131	-0.079	-0.099	0.120	-0.145	0.069	0.186

<i>DoctAdv4</i>	0.066	0.164	0.216	0.428**	0.101	0.135	0.366**	0.132	0.352**	0.198
<i>KnowDG</i>	-0.013	0.007	0.133	0.224**	0.003	0.143	0.226**	0.138	0.229**	0.100
<i>KnowFGP</i>	0.147	0.168	0.308**	0.595**	0.033	0.212	0.455**	0.102	0.354**	0.306**
<i>Know5AD</i>	0.022	0.093	-0.016	0.300**	0.053	-0.047	0.215**	-0.047	0.216**	0.237**
<i>Born2beFat2</i>	-0.160	-0.247**	-0.310**	-0.223**	-0.133	-0.212*	-0.123	-0.065	0.062	0.149
<i>Born2beFat3</i>	-0.065	-0.162	-0.439**	-0.441**	-0.043	-0.350**	-0.345**	-0.244	-0.202	0.033
<i>Born2beFat4</i>	-0.104	-0.124	0.017	-0.270**	-0.036	0.162	-0.125	0.144	-0.114	-0.218*
<i>Born2beFat5</i>	0.259	0.055	0.323	-0.134	-0.254	-0.039	-0.399**	0.185	-0.217*	-0.323**
<i>Pregnant</i>	0.055	-0.445*	0.032	-0.146	-0.465**	-0.015	-0.125	0.280	0.247	-0.037
<i>DocDiabet</i>	0.151	0.256**	0.432**	0.069	0.057	0.269**	-0.072	0.152	-0.146	-0.237*
<i>DiabMedicine</i>	-0.077	0.163	0.280	0.410**	0.302*	0.457*	0.549**	0.207	0.240	0.034
<i>Chronic</i>	0.129	-0.054	0.169	-0.181	-0.156	0.122	-0.291**	0.240	-0.101	-0.263*
<i>Constant</i>	0.347	-0.658**	-0.365	-1.743**	-0.801**	-0.477	-1.710**	0.481	-0.920**	-1.426**

* (**) Statistically significant at the 10% (5%) level.

2. The matching procedure

The next step in the calculation of the propensity score estimator as expressed in equation (8) is the choice of a matching algorithm. Asymptotically, all matching algorithms should yield the same results. However, in practice there are trade offs in terms of bias and efficiency involved with each algorithm. Caliendo and Kopeinig (2005) suggest to try a number of approaches. Hence, we implement six matching algorithms (i.e. one-to-one nearest neighbor, kernel matching, local linear, spline smoothing and radius matching with caliper levels 0.1 and 0.01)³.

Testing the statistical significance of ATT and the computation of standard errors is not a straightforward task because the estimation steps that precede the matching process add variation. We used bootstrapping to address this problem, which we repeated 400 times for each of the matching algorithms to derive the bootstrapped standard errors of ATT. We did not calculate the bootstrap estimator for nearest neighbour matching since Abadie and Imbens (2006) show that the bootstrap variance estimator is invalid for nearest neighbour matching.

Table 3 exhibits the estimated ATT's for each model (standard errors and 95% confidence intervals are also provided). Bold number and asterisks indicate statistically significant effects. We first comment on the statistical significance of our estimates for

³ The matching process was carried out with the *psmatch2* module in Stata (Leuven and Sianesi, 2003).

the unmatched cases. All the models comparing a specific NFP reading level versus the “never” reading NFP case (i.e. 5 vs. 1, 4 vs. 1, 3 vs. 1 and 2 vs. 1) exhibit statistically significant estimates. For example, for (5 vs 1) model in the unmatched case, those who always read the NFP have 0.93 unit higher BMI than those who never read NFPs. Similarly, based on the (2 vs 1) model in the unmatched case, those who rarely read NFPs have 0.70 unit higher BMI than those who never read NFPs. Hence, based on the unmatched cases, one might conclude that nutritional label use increases BMI even though the magnitudes of these effects are quite small. However, after matching, we find that a vast majority of the ATTs are not statistically significant. Hence, in most cases and in general, we cannot reject the null hypothesis of no effect.

To further test the credibility of these results, we conducted robustness tests as well as examined the sensitivity of the results due to unobserved heterogeneity (hidden bias). These are discussed in later sections of the paper.

3. Common support

It is important to check the overlap or common support region for the treated and untreated individuals. A visual analysis of the density distributions of the propensity scores is shown in Figure 1. The bottom-half of each graph shows the propensity score distribution for the non-treated, while the upper-half refers to the treated individuals. Problems would arise if the distributions did not overlap. We imposed the common support using the “minima and maxima comparison”. The basic criterion of this approach is to delete all observations whose propensity score is smaller than the minimum and larger than the maximum in the opposite group. Hence, we removed from our analysis the treated individuals who fall outside the common support region. Table 4 contains the number of observations lost in each model and the propensity score regions after the common support imposition. The number of lost observations in most cases is quite low. Specifically, we lost only a very small fraction (0.5%) of the sample in a vast majority of the models.

Table 3. Average treatment effects on the treated (ATT) for different matching algorithms

	5 vs. 4	5 vs. 3	5 vs. 2	5 vs. 1	4 vs. 3	4 vs. 2	4 vs. 1	3 vs. 2	3 vs. 1	2 vs. 1
	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)	ATT diff. (S.E. ¹)
	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]	[95% CI]
Unmatched	0.208 (0.340)	-0.001 (0.326)	0.232 (0.408)	0.933** (0.323)	-0.208 (0.308)	0.024 (0.386)	0.725** (0.306)	0.232 (0.375)	0.933** (0.289)	0.701* (0.387)
One-to-One nearest neighbor ²	-0.068 (0.480)	-0.631 (0.505)	0.536 (0.688)	0.127 (0.647)	0.337 (0.408)	0.792 (0.667)	0.596 (0.537)	-0.450 (0.554)	0.952** (0.472)	-0.041 (0.687)
Local linear regression	0.076 (0.356) [-0.62, 0.77]	0.055 (0.378) [-0.68, 0.80]	0.311 (0.593) [-0.85, 1.47]	0.661 (0.657) [-0.63, 1.95]	0.020 (0.357) [-0.68, 0.72]	0.445 (0.563) [-0.66, 1.55]	0.642 (0.505) [-0.35, 1.63]	-0.090 (0.458) [-0.99, 0.81]	0.805** (0.367) [0.09, 1.52]	0.712* (0.429) [-0.13, 1.55]
Spline-smoothing	0.111 (0.359) [-0.59, 0.81]	0.052 (0.354) [-0.64, 0.75]	0.254 (0.591) [-0.91, 1.41]	0.671 (0.556) [-0.42, 1.76]	0.035 (0.337) [-0.63, 0.70]	0.326 (0.537) [-0.73, 1.38]	0.671 (0.503) [-0.31, 1.66]	-0.109 (0.455) [-1.00, 0.78]	0.785** (0.359) [0.08, 1.49]	0.648* (0.385) [-0.11, 1.40]
Kernel (epanechnikov)	0.115 (0.380) [-0.63, 0.86]	0.064 (0.362) [-0.65, 0.77]	0.180 (0.579) [-0.95, 1.31]	0.662 (0.642) [-0.60, 1.92]	0.010 (0.331) [-0.64, 0.66]	0.243 (0.532) [-0.80, 1.29]	0.617 (0.466) [-0.30, 1.53]	-0.081 (0.433) [-0.93, 0.77]	0.774** (0.356) [0.07, 1.47]	0.669 (0.431) [-0.18, 1.51]
Radius, Caliper=0.1	0.087 (0.374) [-0.65, 0.82]	0.039 (0.358) [-0.66, 0.74]	0.166 (0.556) [-0.92, 1.26]	0.654 (0.600) [-0.52, 1.83]	-0.055 (0.312) [-0.67, 0.56]	0.023 (0.511) [-0.98, 1.02]	0.676 (0.437) [-0.18, 1.53]	-0.059 (0.396) [-0.84, 0.72]	0.803** (0.359) [0.10, 1.51]	0.680* (0.402) [-0.11, 1.47]
Radius, Caliper=0.01	0.010 (0.377) [-0.73, 0.75]	-0.020 (0.359) [-0.72, 0.68]	0.367 (0.568) [-0.75, 1.48]	0.434 (0.642) [-.82, 1.69]	0.031 (0.346) [-0.65, 0.71]	0.494 (0.504) [-0.49, 1.48]	0.673 (0.472) [-0.25, 1.60]	-0.217 (0.443) [-1.09, 0.65]	0.774** (0.372) [0.04, 1.50]	0.818* (0.431) [-0.03, 1.66]

¹ Bootstrap standard errors for ATT except nearest neighbor, N=400 replications.

² With replacement, no caliper.

* (**) Statistically significant at the 10% (5%) level.

Figure 1. Propensity scores by model

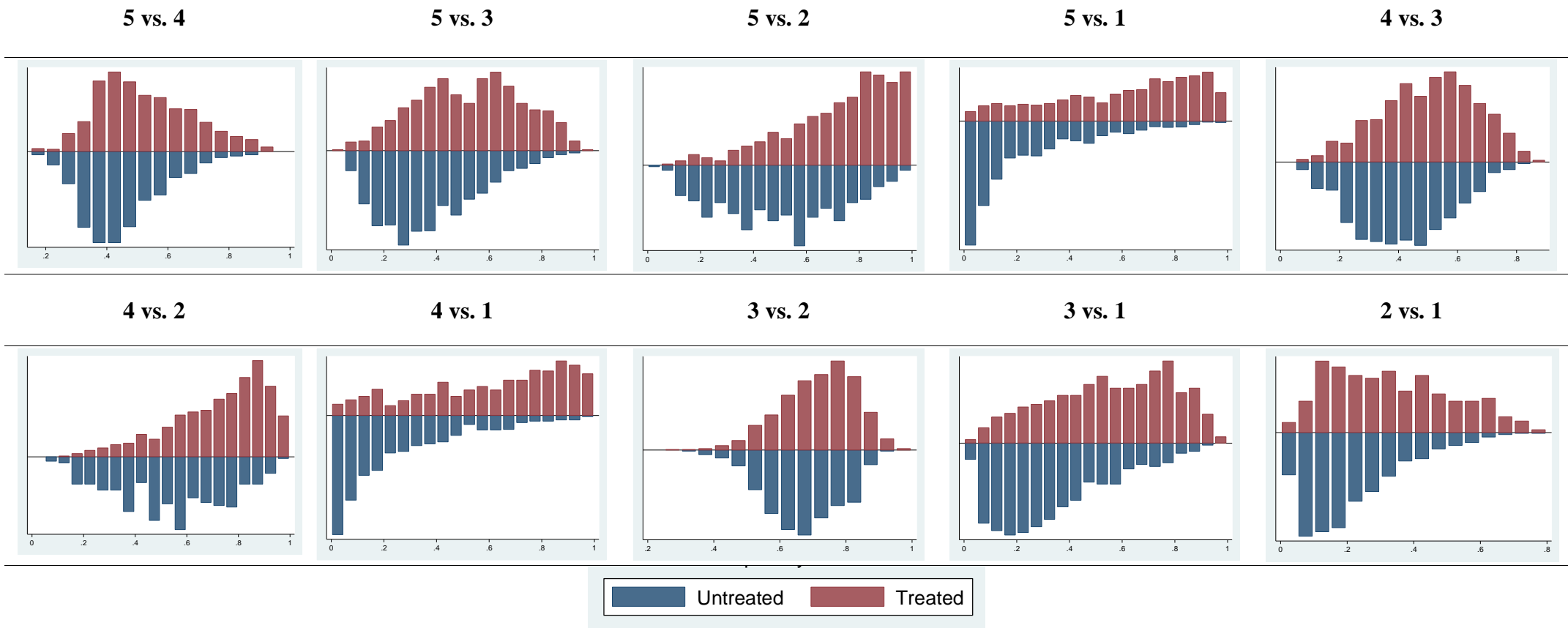


Table 4. Number of treated individuals lost due to common support requirement and range of the propensity scores after comon support imposition^a

<i>Models</i>	Before	After	Lost in %	Probability scores	
	Matching			Min	Max
5 vs. 4	1550	1544	0.39	0.155	0.900
5 vs. 3	1704	1698	0.35	0.056	0.932
5 vs. 2	1163	1144	1.63	0.031	0.984
5 vs. 1	2125	2122	0.14	0.000	0.986
4 vs. 3	1790	1781	0.50	0.076	0.814
4 vs. 2	1249	1209	3.20	0.057	0.960
4 vs. 1	2211	2168	1.94	0.001	0.965
3 vs. 2	1403	1399	0.29	0.328	0.946
3 vs. 1	2365	2335	1.27	0.021	0.917
2 vs. 1	1824	1823	0.05	0.008	0.772

^a We used the minima-maxima restriction as common support condition.

4. Matching quality

In this section, we check whether the matching procedure is able to balance the distribution of the relevant variables. One way to do this is to check if there are differences remaining after conditioning on the propensity score, using the standardized bias (SB) measure proposed by Rubin (1991). For each covariate X , the SB is the difference of the sample means in the treated and matched comparison sub-samples as a percentage of the square root of the average of the sample variances in both groups. For abbreviation, we calculated the means of the SB (MSB) before and after matching by model and matching estimator (Table 5). The overall bias before matching lies between 8.28% and 26.92%. After matching, the bias is significantly reduced for the nearest neighbour, local-linear and spline-smoothing estimators and even more so for the kernel and radius estimators, so that the bias after matching is as low as 2.36% (Radius cal=0.01, “4 vs. 3” model). These results clearly show that the matching procedure is able to balance the characteristics in the treated and the matched comparison groups.

Another approach uses a two sample t-test to check if there are significant differences in covariate means for both groups (see Caliendo and Kopeinig (2008) for a discussion). We performed these tests as well but do not present them due to space considerations. Before matching, several variables exhibit statistically significant differences. However, after matching, the covariates in most cases are balanced and no

significant differences can be found. It appears, however, that the kernel and radius matching estimators are able to more accurately balance the covariates.

We also calculated the pseudo- R^2 before and after matching (see Table 5). The pseudo- R^2 indicates how well the regressors explain the participation probability. After matching, there should be no systematic differences in the distribution of covariates between both groups and the pseudo- R^2 should be low. As the results show, this is true for our matching estimators. Finally, we perform a likelihood ratio test on the joint significance of all regressors. Before matching, the test should be accepted. A rejection of the test after matching reflects a good balancing of the covariates. As exhibited in Table 5, this is also true in most of our cases.

5. *Robustness checks*⁴

To test the robustness of our estimates, we estimate the ATT's on the region of "thick support" defined by $0.33 < \hat{P} X < 0.67$ as suggested by Black and Smith (2004). Black and Smith adopted this approach due to two concerns: (a) the fact that respondents with high estimated propensity scores observed at low levels of treatments may actually represent respondents with measurement error in the treatment variable and (b) residual selection on unobservables, which they demonstrate will have its largest effect on the bias for values of the propensity score in the tails of the distribution. Therefore, they attribute the larger estimates from the thick support area in their study to either (a), (b) or to heterogeneous treatment effects that will have higher impacts for middle values of the propensity score.

As exhibited in Table 6, our thick-support estimates in the majority of the cases are greater than the baseline ones. Hence, similar to Black and Smith, we could not rule out that this difference is due to measurement error in label use or residual selection on unobservables. In the next section, we further test and discuss the effect of any unobserved heterogeneity/hidden bias on our estimates.

⁴ In addition to estimation of the thick support area, we used an older dataset from the 1994-1996 Continuing Survey of Food Intakes for Individuals (CSFII) and performed the matching exercise to this sample as well. Results are generally consistent and supportive of our main finding. In addition, we performed random sub-sampling and out-of-sample predictions with the NHANES dataset and re-estimated ATT's. Main conclusions remain unchanged.

Table 5. Quality of matching indicators

		Models									
		5 vs. 4	5 vs. 3	5 vs. 2	5 vs. 1	4 vs. 3	4 vs. 2	4 vs. 1	3 vs. 2	3 vs. 1	2 vs. 1
<i>Before matching</i>											
	Mean absolute bias	7.28	13.48	18.76	26.02	10.73	16.71	26.92	7.50	18.38	14.02
	Pseudo R ²	0.06	0.11	0.19	0.31	0.07	0.15	0.29	0.04	0.16	0.09
	LR chi ² (p-value)	120.11 (0.00)	261.82 (0.00)	296.43 (0.00)	843.38 (0.00)	165.64 (0.00)	243.92 (0.00)	854.15 (0.00)	69.27 (0.00)	524.08 (0.00)	188.07 (0.00)
<i>After matching</i>											
Nearest Neighbor / Local-linear / Spline- Smoothing	Mean absolute bias	5.06	4.86	8.17	6.29	3.22	6.18	6.40	5.39	4.44	5.08
	Pseudo R ²	0.02	0.02	0.05	0.03	0.01	0.03	0.03	0.03	0.02	0.03
	LR chi ² (p-value)	37.33 (0.45)	44.12 (0.20)	101.20 (0.00)	68.23 (0.01)	25.61 (0.92)	74.15 (0.00)	63.11 (0.01)	69.89 (0.00)	44.91 (0.17)	34.02 (0.61)
Kernel	Mean absolute bias	3.74	3.92	5.79	3.81	2.50	5.88	4.41	4.12	3.04	2.99
	Pseudo R ²	0.011	0.01	0.03	0.02	0.004	0.02	0.01	0.008	0.007	0.009
	LR chi ² (p-value)	22.30 (0.97)	26.24 (0.91)	61.33 (0.01)	30.81 (0.75)	9.41 (1.00)	38.50 (0.40)	29.92 (0.79)	22.22 (0.97)	18.28 (1.00)	10.49 (1.00)
Radius, cal=0.1	Mean absolute bias	3.67	3.71	5.55	4.07	2.62	5.77	4.33	3.75	3.40	3.26
	Pseudo R ²	0.01	0.01	0.03	0.02	0.01	0.02	0.01	0.01	0.01	0.01
	LR chi ² (p-value)	23.64 (0.96)	27.08 (0.88)	60.29 (0.01)	30.95 (0.75)	11.49 (1.00)	36.44 (0.49)	25.45 (0.92)	25.60 (0.92)	19.09 (0.99)	10.56 (1.00)
Radius, cal=0.01	Mean absolute bias	3.51	3.99	6.80	4.08	2.36	6.22	4.08	4.29	3.14	3.19
	Pseudo R ²	0.01	0.01	0.04	0.01	0.005	0.02	0.01	0.011	0.01	0.010
	LR chi ² (p-value)	21.22 (0.98)	25.88 (0.92)	72.07 (0.00)	28.20 (0.85)	11.41 (1.00)	41.19 (0.29)	26.10 (0.91)	28.57 (0.84)	20.83 (0.98)	11.29 (1.00)

Table 6. Average treatment effects on the treated (ATT) for different matching algorithms (thick support region)

	5 vs. 4	5 vs. 3	5 vs. 2	5 vs. 1	4 vs. 3	4 vs. 2	4 vs. 1	3 vs. 2	3 vs. 1	2 vs. 1
	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]	ATT diff. (S.E. ¹) [95% CI]
One-to-One nearest neighbor ²	-0.132 (0.538)	-1.213* (0.697)	0.819 (0.838)	1.267* (0.684)	0.558 (0.469)	1.072 (0.842)	0.420 (0.729)	0.285 (0.696)	1.387** (0.608)	-0.487 (1.028)
Local linear regression	0.001 (0.435) [-0.85, 0.85]	-0.153 (0.544) [-1.21, 0.91]	0.750 (0.730) [-0.68, 2.18]	0.976 (0.613) [-0.23, 2.18]	0.140 (0.380) [-0.60, 0.88]	0.466 (0.666) [-0.84, 1.77]	0.853 (0.665) [-0.45, 2.16]	0.540 (0.525) [-0.49, 1.57]	0.799* (0.465) [-0.11, 1.71]	0.827 (0.722) [-0.59, 2.24]
Spline-smoothing	0.075 (0.428) [-0.76, 0.91]	-0.182 (0.517) [-1.20, 0.83]	0.816 (0.644) [-0.45, 2.07]	0.998* (0.558) [-0.10, 2.09]	0.080 (0.393) [-0.69, 0.85]	0.450 (0.590) [-0.71, 1.61]	1.135* (0.578) [0.002, 2.27]	0.370 (0.535) [-0.68, 1.42]	0.575 (0.443) [-0.29, 1.44]	1.109 (0.688) [-0.24, 2.46]
Kernel (epanechnikov)	0.090 (0.417) [-0.73, 0.91]	-0.236 (0.462) [-1.14, 0.67]	0.724 (0.624) [-0.51, 1.95]	1.117** (0.555) [0.03, 2.20]	0.067 (0.399) [-0.72, 0.85]	0.490 (0.591) [-0.67, 1.65]	1.150* (0.600) [-0.026, 2.33]	0.426 (0.544) [-0.64, 1.49]	0.511 (0.467) [-0.40, 1.43]	1.035 (0.699) [-0.34, 2.41]
Radius, Caliper=0.1	0.037 (0.397) [-0.74, 0.82]	-0.344 (0.457) [-1.24, 0.55]	0.623 (0.606) [-0.56, 1.81]	1.054* (0.555) [-0.03, 2.14]	-0.025 (0.402) [-0.81, 0.76]	0.452 (0.657) [-0.84, 1.74]	1.134* (0.581) [-0.004, 2.27]	0.589 (0.522) [-0.43, 1.61]	0.392 (0.456) [-0.50, 1.29]	1.049 (0.648) [-0.22, 2.32]
Radius, Caliper=0.01	0.074 (0.422) [-0.75, 0.90]	-0.211 (0.547) [-1.28, 0.86]	0.843 (0.657) [-0.44, 2.13]	0.989* (0.586) [-0.16, 2.14]	0.136 (0.410) [-0.67, 0.93]	0.471 (0.670) [-0.84, 1.78]	0.961 (0.671) [-0.35, 2.28]	0.440 (0.530) [-0.60, 1.48]	0.707 (0.444) [-0.16, 1.58]	1.060 (0.744) [-0.40, 2.52]

¹ Bootstrap standard errors for ATT except nearest neighbor, N=400 replications.

² With replacement, no caliper.

* (**) Statistically significant at the 10% (5%) level.

6. Sensitivity analysis for hidden bias/unobserved heterogeneity

Propensity-score matching estimators are based on the assumption that selection is based on observable characteristics. This means that conditional on the observed covariates, the process by which units are selected into treatment is unrelated to unmeasured variables that affect the outcome variable. These estimators are not consistent otherwise. In order to estimate the extent to which such “selection on unobservables” or “hidden bias” may bias the estimates, we conducted a sensitivity analysis which DiPrete and Gangl (2004) call Rosenbaum bounds and is laid out thoroughly in Rosenbaum (2002) and DiPrete and Gangl (2004). This method assesses the sensitivity of significance levels. We emphasize that the method cannot inform us *if* there is unobserved heterogeneity in the data. It can only tell us how much of this unobserved heterogeneity, if any, it would take to change inferences.

When referring to hidden bias, we assume that some characteristics were not controlled for, since these were unobserved, and therefore were not included in X . Therefore, one wants to determine how strongly an unmeasured variable would influence the selection process and undermine the implications of the matching analysis. In brief, this approach assumes that the participation probability π_i is not only determined by observable factors X_i but also by an unobservable component u_i , so that:

$$\pi_i = \Pr(T_i = 1 | X_i) = F(\beta X_i + \gamma u_i) \quad (10)$$

γ is the effect of u_i on the participation decision. If there is no hidden bias γ will be zero. If there is hidden bias, two individuals with the same observed covariates X would have different chances of receiving the treatment. Varying the value of γ allows one to assess the sensitivity of the results with respect to hidden bias and derive bounds of significance levels. The available modules to conduct sensitivity analysis with Rosenbaum bounds can only be implemented in tests for matched (1x1) pairs. Therefore, we conducted sensitivity tests for the one-to-one nearest neighbour and spline smoothing estimators. Tables 13 exhibits the values from Wilcoxon signed rank tests for the average treatment effect on the treated when setting the value of $\Gamma = e^\gamma$ at different levels⁵.

⁵ We used the *rbounds* module in Stata for this estimation (Gangl, 2004).

First, we should describe how Table 7 should be read. For each model and matching estimator, we increased the level of e^γ until the inference about the treatment effect is changed. We report the value of Γ and the critical p-value. The bold cells in the table indicate that these appeared as statistically significant when ATT's were estimated (Table 3). For an ATT that was not statistically significant, the critical value of Γ tells us at which degree of unobserved selection the effect would become significant. For some cells (e.g. "5 vs. 3" model, nearest neighbor) the effect becomes insignificant as the value of Γ is increased. We indicate the 5% level for estimates that turn from insignificant to significant and the 10% level for estimates that turn from significant to insignificant in the sense that these levels represent worst case scenarios.

The opposite applies for the bold cells i.e. we report the value of Γ for which the effect would become insignificant, and in one case the value at which the effect would become significant again. This way we can assess how strong the influence from unobserved variables should be for the estimated ATT to change solely through nonrandom assignment (DiPrete and Gangl, 2004). For example, a critical value for Γ of about 1.20 means that individuals with the same X covariates differ in their odds of participation by a factor of 20%. This result states that the null hypothesis of no treatment effect would not be rejected if an unobserved variable caused the odds ratio of treatment assignment to differ between treatment and comparison groups by 1.20 and if this variable's effect on BMI was so strong as to almost perfectly determine whether the BMI would be bigger for the treatment or the control case in each pair of matched cases in the data.

Based on the results exhibited in Table 7, we can conclude that the nearest neighbor estimator seems to be more sensitive to the existence of unobserved selection than the spline smoothing matching estimator, in the sense that much lower values of Γ are required for an insignificant effect to become significant. Similarly, it would also generally take relatively low values of unobserved selection (between 1.01 to 1.20) to change a statistically significant effect into a statistically insignificant effect with the nearest neighbor estimator. For the spline smoothing matching estimator, our sensitivity analysis suggests that it would take much higher values of Γ to change an insignificant effect into a significant effect. Given these results, it would be more prudent than in our case to rely more on the spline-smoothing estimates than the nearest

neighbor estimates. Given that a vast majority of our spline-smoothing ATT estimates are statistically insignificant (i.e., 8 out of 10), our sensitivity analysis suggests that these estimates would remain statistically insignificant even if we had substantial unobserved heterogeneity/selection. In other words, it is not likely that nutritional label use will have an effect on BMI even in the presence of unobserved heterogeneity.

Table 7. Rosenbaum bounds for BMI treatment effects

<i>Models</i>	One-to-One nearest neighbor		Spline-smoothing	
	Gamma	p-critical	Gamma	p-critical
<i>5 vs. 4</i>	1.12	<i>0.037</i>	1.36	<i>0.046</i>
<i>5 vs. 3</i>	1.01 1.06	<i>0.031</i> <i>0.118</i>	1.42	<i>0.033</i>
<i>5 vs. 2</i>	1.04	<i>0.033</i>	1.34	<i>0.043</i>
<i>5 vs. 1</i>	1.06	<i>0.039</i>	1.14	<i>0.036</i>
<i>4 vs. 3</i>	1.08	<i>0.038</i>	1.42	<i>0.045</i>
<i>4 vs. 2</i>	1.01 1.14	<i>0.002</i> <i>0.107</i>	1.28	<i>0.042</i>
<i>4 vs. 1</i>	1.01 1.16	<i>0.002</i> <i>0.141</i>	1.12	<i>0.043</i>
<i>3 vs. 2</i>	1.06	<i>0.031</i>	1.01 1.14	<i>0.002</i> <i>0.123</i>
<i>3 vs. 1</i>	1.20	<i>0.101</i>	1.02	<i>0.127</i>
<i>2 vs. 1</i>	1.38	<i>0.050</i>	1.01 1.18	<i>0.431</i> <i>0.048</i>

V. DISCUSSION AND CONCLUSIONS

So much attention has been given lately to the issue of nutritional labeling due to the obesity problem. The hypothesis is that nutritional label use can reduce obesity rates. Previous studies, as discussed earlier, have generally found that nutritional label use can improve dietary outcomes. However, it is unknown and unclear if nutritional label use can indeed influence body weight outcomes. This is the simple aim of our study. Using propensity matching technique, our results generally suggest that nutritional label use does not have an effect on BMI.

The FAFH sector is under increasing pressure to provide nutritional information in restaurants and fast food places. Much of the arguments in favor of a mandatory nutritional labeling law in the FAFH sector has stemmed from the supposedly beneficiary impact of nutrition information in the Food-At-Home market. The New York City Board of Health has already taken one step forward by requiring the city's restaurant chains to show calorie information on their menus and menu board. The new regulation came into effect in April 2008 and applies to any chain restaurant in New York City that has 15 or more outlets in the US. One of the benefits of this law was estimated to be the reduction in the number of obese New Yorkers by 150,000 over the next five years. Given our finding, this projection might be overstated.

Since the NLEA is only for the food at home market, it is not clear either if mandatory nutritional labelling in the FAFH market is warranted given our findings. More research is needed to specifically analyze the effects of nutritional label use in the FAFH market on body weight and other health outcomes. Unfortunately, we do not currently know of any existing comprehensive datasets that would enable researchers to conduct such analysis at the moment. Future studies should also attempt to definitively assess the possible reasons on why reading nutritional labels would not reduce BMI. One possible explanation that could be evaluated is the remedy message explanation, a phenomenon well founded in the marketing literature. Nutritional labeling can be seen as a disclosure remedy, that has the aim to correct market failure related to the inadequate provision of information (Seiders and Petty, 2004). Ironically, remedy messages boomerang on the people who are intended to be helped the most (Bolton et al., 2006) because some consumers appear less risk averse when remedies are available. For example, in an experiment, Bolton et al. (2006) found that a remedy message for a

fat-fighting pill undermined food fat content perceptions and increased high-fat eating intentions as problem status (concerns about body image) increased. Another possible reason is moral hazard since it is possible that individuals who read nutritional labels take less precaution in other areas of weight control.

VI. REFERENCES

- Abadie, Alberto and Imbens, Guido W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74: 235-267.
- Becker, Sascha O. and Ichino, Andrea (2002). Estimation of average treatment effects based on propensity scores. *The Stata Journal*, 2: 358-377.
- Black, Dan A. and Smith, Jeffrey A. (2004). How robust is the evidence on the effects of college quality? Evidence from matching. *Journal of Econometrics*, 121: 99-124.
- Bolton, Lisa E., Cohen, Joel B. and Bloom, P. N. (2006). Does marketing products as remedies create "Get out of jail free cards"? *Journal of Consumer Research*, 33: 71-81.
- Caliendo, Marco and Kopeinig, Sabine (2005). Some practical guidance for the implementation of propensity score matching. *Discussion Paper 485, DIW German Institute for Economic Research: Berlin*.
- Caliendo, Marco and Kopeinig, Sabine (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys*, 22: 31-72.
- Coulson, Neil S. (2000). An application of the stages of change model to consumer use of food labels. *British Food Journal*, 102: 661-668.
- DiPrete, Thomas A. and Gangl, Markus (2004). Assessing bias in the estimation of causal effects: Rosenbaum bounds on matching estimators and instrumental variables estimation with imperfect instruments. *Sociological Methodology*, 34: 271-310.
- European Advisory Services (2004). The introduction of mandatory nutrition labelling in the European Union. Brussels: European Advisory Services.
- Gangl, Markus (2004). Stata module to perform Rosenbaum sensitivity analysis for average treatment effects on the treated.
- Guthrie, Joanne F., Fox, Jonathan J., Cleveland, Linda E. and Welsh, Susan (1995). Who uses nutritional labeling, and what effects does label use have on diet quality? *Journal of Nutrition Education*, 27: 163-172.
- Heckman, James J., Ichimura, Hidehiko and Todd, Petra E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The Review of Economic Studies*, 64: 605-654.
- Heckman, James J., Ichimura, Hidehiko and Todd, Petra E. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65: 261-294.
- Heckman, James J. and Navarro-Lozano, Salvador (2004). Using matching, instrumental variables, and control functions to estimate economic choice models. *The Review of Economics and Statistics*, 86: 30-57.
- Hirano, Keisuke and Imbens, Guido W. (2004). The propensity score with continuous treatments. in: A. Gelman and X. Meng (eds.) *Missing data and Bayesian methods in practice: Contributions by Donald Rubin's statistical family*. New York: Wiley.
- Imbens, Guido W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika*, 87: 706-710.
- Imbens, Guido W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics*, 86: 4-29.
- Joffe, Marshall M. and Rosenbaum, Paul R. (1999). Invited commentary: Propensity scores. *American Journal of Epidemiology*, 150: 327-333.

- Kim, Sung-Yong, Nayga, Rodolfo M., Jr. and Capps, Oral, Jr. (2000). The effect of food label use on nutrient intakes: An endogenous switching regression analysis. *Journal of Agricultural and Resource Economics*, 25: 215-231.
- Kim, Sung-Yong, Nayga, Rodolfo M., Jr. and Capps, Oral, Jr. (2001). Food label use, self-selectivity, and diet quality. *The Journal of Consumer Affairs*, 35: 346-363.
- Lechner, Michael (2002). Program heterogeneity and propensity score matching: An application to the evaluation of active labor market policies. *The Review of Economics and Statistics*, 84: 205-220.
- Lechner, Michael (2007). A note on endogenous control variables in causal studies. *Statistics & Probability Letters*, 78: 190-195.
- Leuven, E. and Sianesi, b. (2003). PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing.
- Lin, Chung-Tung Jordan and Lee, Jonq-Ying (2003). Dietary fat intake and search for fat information on food labels: New evidence. *Consumer Interests Annual*, 49.
- Loureiro, Maria L. and Nayga, Rodolfo M., Jr. (2005). International dimensions of obesity and overweight related problems: An economics perspective. *American Journal of Agricultural Economics*, 87: 1147-1153.
- Lu, Bo, Zanutto, Elaine, Hornik, Robert and Rosenbaum, Paul R. (2001). Matching with doses in an observational study of a media campaign against drug abuse. *Journal of the American Statistical Association*, 96: 1245-1253.
- Neuhouser, Marian L., Kristal, Alan R. and Patterson, Ruth E. (1999). Use of food nutrition labels is associated with lower fat intake. *Journal of the American Dietetic Association*, 99: 45-53.
- Rosenbaum, Paul R. (2002). *Observational studies*, New York: Springer-Verlag.
- Rosenbaum, Paul R. and Rubin, Donald B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70: 41-56.
- Rubin, Donald B. (1991). Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47: 1213-1234.
- Russo, J. Edward, Staelin, Richard, Nolan, Catherine A., Rusell, Gary J. and Metcalf, Barbara L. (1986). Nutrition information in the supermarket. *Journal of Consumer Research*, 13: 48-70.
- Seiders, Kathleen and Petty, Ross D. (2004). Obesity and the role of food marketing: A policy analysis of issues and remedies. *Journal of Public Policy & Marketing*, 23: 153-169.
- Teisl, Mario F., Bockstael, Nancy E. and Levy, Alan S. (2001). Measuring the welfare effects of nutrition information. *American Journal of Agricultural Economics*, 83: 133-149.
- Variyam, Jayachandran N. (2004). Evaluating the effect of nutrition labels: a quasi-experimental approach. *Creating and Using Evidence in Public Policy Analysis and Management*. Atlanta, GA, USA: Twenty-Sixth Annual APPAM Research Conference.
- Variyam, Jayachandran N. (2008). Do nutrition labels improve dietary outcomes? *Health Economics*, 17: 695-708.