

RESEARCH ARTICLE

Evaluating Propensity Score Methods in a Quasi-Experimental Study of the Impact of Menu-Labeling

Stephanie L. Mayne*, Brian K. Lee, Amy H. Auchincloss

Department of Epidemiology and Biostatistics, School of Public Health, Drexel University, Philadelphia, Pennsylvania, United States of America

* slm387@drexel.edu



Abstract

OPEN ACCESS

Citation: Mayne SL, Lee BK, Auchincloss AH (2015) Evaluating Propensity Score Methods in a Quasi-Experimental Study of the Impact of Menu-Labeling. PLoS ONE 10(12): e0144962. doi:10.1371/journal.pone.0144962

Editor: Marc A Adams, Arizona State University, UNITED STATES

Received: June 10, 2015

Accepted: November 25, 2015

Published: December 17, 2015

Copyright: © 2015 Mayne et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: This research is with human subjects and thus release of the data is subject to ethics review board review following HHS-National Institute of Health guidelines. To obtain access to the data, please contact: Amy Auchincloss, Associate Professor Dept of Epidemiology and Biostatistics, School of Public Health, Drexel University: aha27@drexel.edu; and Danyelle Gibson, Assistant Director, Human Research Protection, Office of Research, Drexel University: Danyelle.S.Gibson@drexel.edu.

Funding: The authors have no support or funding to report.

Background

Quasi-experimental studies of menu labeling have found mixed results for improving diet. Differences between experimental groups can hinder interpretation. Propensity scores are an increasingly common method to improve covariate balance, but multiple methods exist and the improvements associated with each method have rarely been compared. In this re-analysis of the impact of menu labeling, we compare multiple propensity score methods to determine which methods optimize balance between experimental groups.

Methods

Study participants included adult customers who visited full-service restaurants with menu labeling (treatment) and without (control). We compared the balance between treatment groups obtained by four propensity score methods: 1) 1:1 nearest neighbor matching (NN), 2) augmented 1:1 NN (using caliper of 0.2 and an exact match on an imbalanced covariate), 3) full matching, and 4) inverse probability weighting (IPW). We then evaluated the treatment effect on differences in nutrients purchased across the different methods.

Results

1:1 NN resulted in worse balance than the original unmatched sample (average standardized absolute mean distance [ASAM]: 0.185 compared to 0.171). Augmented 1:1 NN improved balance (ASAM: 0.038) but resulted in a large reduction in sample size. Full matching and IPW improved balance over the unmatched sample without a reduction in sample size (ASAM: 0.049 and 0.031, respectively). Menu labeling was associated with decreased calories, fat, sodium and carbohydrates in the unmatched analysis. Results were qualitatively similar in the propensity score matched/weighted models.

Competing Interests: The authors have declared that no competing interests exist.

Conclusions

While propensity scores offer an increasingly popular tool to improve causal inference, choosing the correct method can be challenging. Our results emphasize the benefit of examining multiple methods to ensure results are consistent, and considering approaches beyond the most popular method of 1:1 NN matching.

Introduction

Changes to diet, particularly an increase in energy intake, have been implicated as a major cause of the rising obesity epidemic.[1, 2] Food prepared away from home is typically higher in calories and lower in nutrients than food prepared at home, yet comprises approximately one-third of all calories purchased in the United States.[3] As of 2008, many municipal and state governments have mandated that chain restaurants display nutrition information at point of purchase. The regulations aim to promote nutrition awareness which may increase healthier offerings and choices among restaurants and customers, respectively. However, quasi-experimental studies evaluating the effectiveness of these policies have reported mixed results.[4–9] One of the few studies to find a labeling impact was Auchincloss et al (2013) which studied the impact of the city of Philadelphia’s menu-labeling ordinance.[10] The study found that customers at full-service restaurants with labeling purchased fewer calories, sodium, and saturated fat compared to customers at restaurants without labeling.[5] However, the cross-sectional study design was suboptimal and there was imperfect covariate balance between customers who dined at labeled versus unlabeled restaurants. Imbalance can lead to selection bias, in which those who have the opportunity to be exposed to the policy differ from those who do not on certain measured or unmeasured characteristics related to the outcome of interest.

In this study, we re-analyze data from Auchincloss et al (2013) via propensity score methods and compare the effect of menu labeling on nutritional outcomes, including purchased calories, saturated fat, sodium, and carbohydrates. The objectives of the study were to determine which method resulted in optimal balance between study arms and to determine if the association between menu labeling and nutritional outcomes was similar across methods.

Matching or weighting on propensity scores, defined as the odds or probability of receiving treatment based on a set of covariates,[11] are commonly used methods to reduce the imbalance of covariates between treatment groups. Propensity scores provide a balancing measure that can be used to match treated and untreated individuals based on their probability of being exposed to the intervention, and can also be used as weights or to stratify the study sample. Propensity score matching and weighting have been proposed as alternatives to multivariable regression models to address imbalance between groups in natural experiments,[12] and these approaches have been used in several previously published studies of responses to policy or built environment changes using quasi-experimental designs.[13–15] However, the relative performance of the different methods available has infrequently been compared, in settings that may not generalize to all observational study contexts.[16, 17]

Methods

Sampling

This study is a secondary analysis of a previously published study.[5] One large, mid-priced national chain restaurant was selected for inclusion in the study because it had a sufficient

number of outlets in Philadelphia, where menu labeling was mandated, and outside Philadelphia, where no menu labeling was required. A convenience sample was included with two outlets in Philadelphia (case restaurants) and five outlets from within a 130-mile radius outside of the city (control restaurants). Menus were identical between treatment and comparison restaurants except that at case restaurants, nutrition information was clearly displayed in labels next to all food items, which included calories, grams of saturated fat, trans fat, carbohydrates, and milligrams of sodium.

Data collection

Data sources included receipts and surveys collected from customers exiting the restaurants between 6-9pm on Sundays and Tuesdays-Thursdays over a one-month period in August, 2011. All data were anonymous with no identifiers. Surveys were used to obtain information on whether respondents saw nutrition information and whether this information affected what they ordered, demographic characteristics, frequency of dining at chain restaurants, and whether a health professional had recommended limiting calories, fat, carbohydrates and sodium. Study staff recorded participants' body size (thin/average, overweight, and severely overweight) and whether the party included children.

Outcome

The primary outcomes were number of calories purchased from food and beverages combined, number of food calories, grams of saturated fat, grams of carbohydrates, and milligrams of sodium in food purchases.

Covariates

Covariates included age, gender, race, education, income, body size, whether a health professional cautioned the participant about his/her diet, frequency of dining at chain restaurants, day of the week, whether there were children in the party, and whether the customer made any substitutions or customizations to their order. Day of the week, the presence of children in the party, and whether substitutions were made were included as selection factors that were relevant to the outcomes of interest.

Covariates were slightly different than those used in the original paper in order to include additional potential confounders. [5]

Statistical Analysis

Descriptive statistics were calculated to compare the covariate balance between customers at restaurants with and without labeling. Bivariate comparisons were made using chi-squared and t-tests. Multivariable linear regression models were used to examine mean differences in dietary outcomes between participants at labeled (treatment) and unlabeled (control) restaurants, controlling for the covariates listed above. Five separate models were used, one for each of the outcomes of interest. These results were compared to results from the propensity score methods described below.

Propensity scores. Propensity scores were calculated using a logistic regression model, with the outcome of patronizing a restaurant with labeling versus without. The covariates listed above were included as predictors. The resulting propensity score represented each participant's predicted probability of patronizing a labeled restaurant based on the set of observed covariates. Matching was performed using the raw propensity scores, rather than using a logit

transformation as has been used in some previous studies [18, 19], as this approach produced better covariate balance.

Propensity score analyses generally estimate either the average treatment effect (ATE) which represents the average effect of treatment on the whole study population; or the average treatment effect on the treated (ATT) which represents the average effect of treatment on those subjects who actually attended restaurants with menu labeling (the treated group). The estimand of interest in the present study was the ATT because in this study, as in many observational studies, the treated persons differed from untreated persons in significant ways. As such, estimating the ATT is more appropriate in this case.

Propensity Score Methods. After deriving each participant's propensity score, the scores were used to match or weight treated and control individuals in preparation for the final outcome model. Four different methods were used. Matching was implemented using the MatchIt package in R (Methods 1 and 2) [20] and Matchit with the Optmatch package integrated (Method 3).[20, 21]

For Method 1 we selected the method most commonly used in studies of propensity scores, [16] greedy 1:1 nearest neighbor matching. This approach selects, for each individual in the treated group, the individual in the control group with the closest propensity score and matches one pair at a time[22]. "Greedy" matching indicates that when two treated individuals have the same control individual with the closest propensity score, the control will be matched to whichever treatment individual comes first in the matching order.[23] Once a match has been assigned, matches are not broken or revisited by the algorithm. Individuals in the control group who are not selected as a match for a treated individual are dropped from the matched dataset.

Method 2 augmented the 1:1 nearest neighbor matching with an additional exact match on race, a confounder that remained fairly imbalanced after matching, and a caliper of 0.2 standard deviations of the propensity score to ensure a closer match between treatment and control individuals. The caliper set a limit such that matched treated and control individuals' propensity scores could not differ by greater than 0.2 standard deviations. A caliper of 0.2 has been previously described as an optimal width. [24] Due to the additional criteria placed on the match, this method would retain only about 70% of the study sample.

Method 3 was full matching,[18, 19, 23, 25] a lesser used approach that makes use of all observations in the data and has been found effective at reducing bias through producing very good covariate balance between treatment and comparison groups.[19] Full matching forms a series of matched subclasses where each set contains either at least one treated individual and multiple controls, or at least one control and multiple treated individuals. Observations are matched by minimizing a weighted average of the estimated distance measure between each treated subject and each control. In order to improve precision of the estimate and reduce dependence on any particular propensity score specification,[18, 26] we constrained the subclass ratios of treated to controls to be no less than 0.5 and no more than 2.0. This meant that within each matched subclass, the ratio of treated to controls was not fixed a priori but nevertheless would not vary widely. The range of 0.5 to 2.0 was chosen based on prior literature[18, 19] [the original sample ratio was approximately 1:1 (327 treated/321 control individuals)].

Finally, Method 4 was inverse probability weighting (IPW) by the propensity score.[16] IPW makes use of all observations in the dataset, is a widely used approach, and has been shown to provide good balance between groups.[27] In order to provide an appropriate comparison to the results obtained by matching methods, which inherently estimate the ATT rather than the ATE, we calculated each individual's odds of treatment as weights in order to estimate the ATT.

Thus, individuals in the treatment group were assigned weights equal to 1, and individuals in the control group were assigned weights equal to $[\text{propensity score}]/[1-\text{propensity score}]$. [28] This enables us to compare results from IPW to those obtained by matching methods, since by matching the control group to the treated group based on propensity scores, these methods inherently estimate the ATT rather than the ATE.

In order to evaluate the effectiveness of each method, balance was examined between the treated and untreated groups by calculating the standardized bias for each covariate in the propensity score model. The standardized bias is calculated by taking the difference in means for a given covariate between the treatment and control groups and dividing by the standard deviation in the treatment group. Standardized biases of less than 0.25 suggest good balance between the groups. [23] Boxplots were then constructed to display the absolute values of standardized biases before and after matching or weighting. [19]

After calculating the individual standardized biases for each covariate, the average standardized absolute mean distance (ASAM) was calculated for each method to provide a global comparison of the balance across methods. The ASAM is the absolute value of the standardized difference of means between the treatment and control groups (standardized by the standard deviation of each covariate in the treatment group), averaged across all of the covariates. A smaller ASAM indicates better balance between groups.

Determining the Effect of Menu Labeling on Nutritional Outcomes After Matching and Weighting. In order to determine whether measures of association between menu labeling and nutritional outcomes varied based on the propensity score method used, final outcome regression models were implemented using the matched or weighted data. For the nearest-neighbor methods (Methods 1 and 2), unweighted regression models were implemented using the reduced matched data set, adjusting for all covariates in the model in order to perform a doubly robust regression. [29] This is the same analysis as using the original sample but using the matched sample instead.

For the full matching analysis (Method 3), an ATT weighting approach was applied to the regression models in the full dataset. Within each matched set, each treated individual received a weight of 1 and controls received a weight proportional to the number of treated divided by the number of controls in that set. [23] As treated individuals receive a weight of 1, and control individuals are weighted to the treated individuals, this approach estimates the ATT. These weights were applied to linear regression outcome models that also controlled for the covariates of interest. For Method 4, the inverse probability weights (1 for the treated group, $[\text{propensity score}]/[1-\text{propensity score}]$ for the control group) were applied to adjusted linear regression models.

All analyses were conducted using R version 3.0.3. [30] No identifiers were collected in the original study. The Drexel University Institutional Review Board approved the original study and determined that this retrospective secondary analysis of anonymous data did not qualify as human subjects research.

Results

In total, 648 participants completed the survey (327 at restaurants with labeling, 321 at restaurants without labeling). Table 1 describes the balance of covariates between the two treatment groups. Customers at labeled restaurants were younger on average (mean 35 years old compared to 38 at unlabeled restaurants), a larger proportion was non-white, and a smaller proportion had income \geq \$60,000. Customers at labeled sites were also less likely to dine out at least once per week, and to be enrolled on Sundays.

Table 2 presents the covariate balance between participants at labeled and unlabeled restaurants before and after implementing the four propensity score methods. The standardized

Table 1. Initial Covariate Balance between Treated (Labeled) and Control (Unlabeled) Restaurant Customers.

Variable	Labeled Restaurants N (%)	Unlabeled Restaurants N (%)	p-value ¹
N	327	321	
Age (mean, SD)	35.1 (13.0)	38.4 (14.0)	0.002
Sex			
Female	192 (59%)	199 (62%)	0.4
Male	135 (41%)	122 (38%)	
Race/Ethnicity			
White	100 (30%)	153 (48%)	<0.001
Black	182 (56%)	142 (44%)	
Hispanic and Other	45 (14%)	26 (8%)	
Income			
<\$35,000	92 (28%)	65 (20%)	0.004
\$35–60,000	101 (31%)	84 (26%)	
> = \$60,000	134 (41%)	172 (54%)	
Education			
High School or Less	78 (24%)	79 (25%)	0.8
Technical or Associate's Degree	51 (15%)	58 (18%)	
Bachelor's Degree	133 (41%)	121 (38%)	
Graduate School	64 (20%)	63 (19%)	
Cautioned about Diet by Health Professional	70 (21%)	66 (21%)	0.8
Body Size			
Not severely overweight	291 (89%)	276 (86%)	0.3
Severely Overweight	36 (11%)	45 (14%)	
Frequency of Dining at Chain Restaurants			
> = Once per Week	123 (38%)	144 (45%)	0.06
<Once per Week	204 (62%)	177 (55%)	
Children in Party			
Yes	80 (24%)	75 (23%)	0.08
No	140 (43%)	184 (57%)	
Missing	107 (33%)	62 (20%)	
Customized Order	58 (18%)	64 (20%)	0.5
Day of the Week			
Sunday	16 (5%)	55 (17%)	<0.001
Tuesday	120 (37%)	74 (23%)	
Wednesday	92 (28%)	11 (35%)	
Thursday	99 (30%)	81 (25%)	

¹P-values calculated by chi-squared tests

doi:10.1371/journal.pone.0144962.t001

biases prior to matching reflect the imbalances shown in [Table 1](#), with large standardized biases for age, race, income, day of the week, and number of children in the party. Method 1, greedy 1:1 nearest neighbor matching that was identical to the original sample except for 6 treated individuals who were not matched, resulted in worse overall balance than the original sample. Method 2, augmented 1:1 nearest neighbor matching, improved the balance between groups for most covariates, but resulted in the removal of 94 treated and 88 control individuals who did not have appropriately similar matches. Full matching (Method 3) and IPW (Method 4) resulted in reductions in the standardized biases for most covariates. Methods 2, 3 and 4 reduced the ASAM from the pre-matching level of 0.171, reflecting improved covariate

Table 2. Comparison of Absolute Standardized Biases^a of Covariates between Treated (Labeled) and Control (Unlabeled) Subjects after Propensity Score Matching and Weighting^b.

Variable	Unmatched Sample	Method 1. Greedy 1:1 Nearest Neighbor Matching	Method 2. Greedy Nearest Neighbor + Exact Match + Caliper of 0.2 SD	Method 3. Full Matching, Constrained	Method 4. Inverse Probability Weighting ^d
Age	0.253*	0.266*	0.012	0.072	0.010
Sex	0.067	0.076	0.009	0.022	0.011
Race	0.353*	0.372*	0.000	0.076	0.029
Income	0.249	0.269*	0.026	0.074	0.042
Education	0.040	0.050	0.004	0.013	0.014
Body Size	0.096	0.119	0.027	0.083	0.059
Cautioned about Diet	0.021	0.030	0.052	0.030	0.047
Frequency of dining out	0.149	0.173	0.142	0.032	0.011
Day of the week	0.272*	0.271*	0.004	0.010	0.053
Customizations	0.056	0.073	0.011	0.024	0.034
Children in Party	0.323*	0.335*	0.129	0.101	0.026
Average Standardized Absolute Mean Difference (ASAM) ^c	0.171	0.185	0.038	0.049	0.031
N Treated/N Discarded	327	321/6	233/94	327/0	327/0
N Control/N Discarded	321	321/0	233/88	321/0	321/0

^aAbsolute standardized bias is the absolute value of the weighted difference in means between the treatment and control group divided by the standard deviation in the treatment (labeled) group.

^bPropensity scores were calculated using logistic regression.

^cThe ASAM is calculated by taking the average of the absolute values of the standardized biases for all covariates used to calculate the propensity score (not including the propensity score itself).

^dInverse probability weights indicate odds of being in the treatment (labeled) group and were calculated as follows: 1 for treated, [propensity score]/[1-propensity score] for control

*indicates standardized bias >0.25

doi:10.1371/journal.pone.0144962.t002

balance. [Fig 1](#) presents boxplots of the absolute values of the standardized biases for the covariates of interest, before and after implementing the four methods. Consistent with the results in [Table 2](#), Methods 2, 3, and 4 resulted in absolute standardized biases that were closest to zero. [Fig 1](#) and [Table 2](#) together indicate that Method 4 (IPW) had the best bias reduction and uniformly good balance across all covariates.

The results of the linear regression models assessing the association of menu labeling with calories, saturated fat, carbohydrates, and sodium purchased, before and after matching, are presented in [Table 3](#). In the pre-match, unweighted models, menu labeling was associated with a statistically significant reduction in total calories purchased (-179.6), food calories (-178.2), saturated fat (-46), carbohydrates (-16.1), and sodium (-279.4). In models accounting for each of the four propensity score methods, results were substantively similar but slightly attenuated for most outcomes. Specifically, menu labeling was associated with a reduction in total calories of -149.0 and -135.3 for constrained full matching and IPW, respectively, although these reductions remained statistically significant. The same pattern was observed for food calories (-149.7 and -139.1 for full matching and IPW). The only outcome that was no longer significantly associated with menu labeling after using a propensity score-based method was saturated fat using Method 2 (nearest neighbor matching with a caliper and exact match on race), which was -3.3 (-7.6, 0.9). All associations were in the same direction as in the initial models.

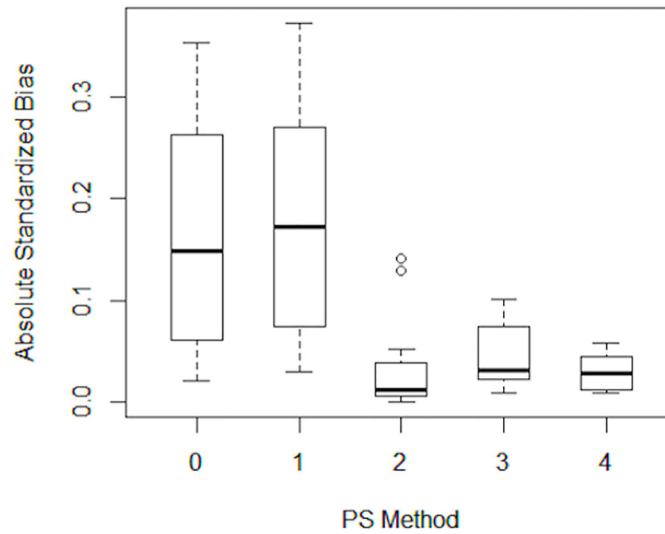


Fig 1. Boxplots of Absolute Standardized Biases for the Covariates in the Propensity Score Model. Methods. 0-original sample, no matching or weighting; 1- greedy 1:1 nearest neighbor matching, 2- greedy 1:1 nearest neighbor matching with exact match on race and caliper of 0.2, 3- full matching (constrained), 4- inverse probability weighting

doi:10.1371/journal.pone.0144962.g001

Discussion

In this study of a menu labeling natural experiment, the initial imbalance in characteristics between customers at labeled and unlabeled restaurants was improved by three out of four propensity score matching methods. Of these methods, constrained full matching and IPW

Table 3. Effect of Menu Labeling on Nutritional Outcomes^a Before and After Propensity Score Matching/Weighting^b.

	Original Adjusted Regression Models	Method 1. Greedy 1:1 Nearest Neighbor Matching	Method 2. Greedy Nearest Neighbor + Exact Match on Race + Caliper of 0.2 SD	Method 3. Full Matching, Constrained	Method 4. Inverse Probability Weighting ^c
Total Calories (kcal)	-179.6 (-310.0, -49.2)*	-165.3 (-297.8, -32.8)*	-167.7 (-314.3, -21.2)*	-149.0 (-273.4, -24.6)*	-135.3 (-260.0, -10.6)*
Food Calories (kcal)	-178.2 (-299.3, -57.1)*	-161.4 (-284.3, -38.4)*	-167.9 (-305.8, -30.1)*	-149.7 (-264.4, -34.9)*	-139.1 (-255.6, -22.5)*
Food Saturated Fat (g)	-4.6 (-8.4, -0.9)*	-4.3 (-8.0, -0.5)*	-3.3 (-7.6, 0.9)	-3.6 (-7.1, 0.0)*	-4.5 (-8.2, -0.7)*
Food Carbohydrates (g)	-16.1 (-27.5, -4.7)*	-14.8 (-26.4, -3.3)*	-15.8 (-29.0, -2.6)*	-16.6 (-27.5, -5.6)*	-14.7 (-25.7, -3.8)*
Food Sodium (mg)	-279.4 (-515.5, -43.4)*	-252.5 (-492.0, -12.9)*	-338.2 (-608.8, -67.5)*	-282.5 (-512.9, -52.1)	-212.4 (-440.4, 15.7)

^aEach model presents the mean difference in the nutrition outcome in question between customers at labeled restaurants and customers at unlabeled restaurants, adjusted for the following covariates: age, sex, race/ethnicity, income, education, body size, whether they were cautioned about their diet, frequency of dining out, day of the week, whether there were children in the party, and whether the order was customized.

^bResults reported here differ from regression models reported in Auchincloss et al (2013) due to (a) adding three additional covariates (whether a health professional cautioned the participant about his/her diet, whether there were children in the party, and whether the customer made any substitutions or customizations to their order) as covariates to the unweighted model and (b) estimating the ATT in the Inverse Probability weighted regression models rather than the ATE. The original paper reported that the treatment difference in overall calories was -155.0 (-284.0, -27.0) in the unweighted model and -166.6 (-286.7, -46.5) in the weighted model.

*indicates p<0.05

^cInverse probability weights indicate odds of being in the treatment (labeled) group and were calculated as follows: 1 for treated, [propensity score]/[1-propensity score] for control

doi:10.1371/journal.pone.0144962.t003

resulted in a substantial reduction in standardized bias without a reduction in sample size. The association between menu labeling and nutritional outcomes was fairly stable across methods and was slightly attenuated, though substantively similar, compared to the initial unmatched/unweighted regression models. These findings indicate that good balance can be achieved using several methods. In addition, our findings strengthen inferences about the association between menu labeling at sit-down restaurants and reduction in calories, saturated fat, carbohydrates, and sodium purchased, as the results obtained after making the treatment and control group more comparable were substantively similar to those observed in the original analysis published from this study.[\[5\]](#)

Researchers are increasingly turning to propensity scores as a tool to improve causal inference. However, many methods are available and choosing the right method can be challenging. 1:1 nearest neighbor matching on the propensity score is one of the most common methods; [\[26\]](#) however, we found that this form of matching did not improve the standardized bias in this study, as the matched sample was very similar to the original sample and no weights were applied. Augmenting the match with a caliper, a method that has been suggested previously to improve the quality of the matching process,[\[16\]](#) and an exact match on an unbalanced covariate resulted in a very close match but substantially reduced the sample size, an important consideration to keep in mind when implementing these methods. Loss of power is a possible limitation of imposing such strict criteria on the matching, as only ~70% of the study population was retained. In addition, 94 treated individuals were discarded. Loss of treated individuals may shift the estimation target as the treatment effect may no longer represent the effect for all participants in the treatment group.[\[31\]](#)

In consideration of these limitations, our findings suggest that researchers may benefit from several lesser-used but easy-to-implement methods, namely full matching and IPW. Full matching, which has been described as a compromise between k:1 matching and subclassification,[\[19\]](#) is more flexible than nearest neighbor matching because the number of comparison individuals matched to each treated individual can vary depending on how many good matches are actually available. Full matching also makes use of all available data (does not drop any individuals from the analysis), and has been found to perform better than k:1 matching in simulated and real data.[\[19, 32\]](#) Our finding that constrained full matching resulted in an overall reduction in standardized bias is consistent with a prior study comparing various methods of matching in a cross-sectional study of the relationship between adolescent marijuana use and adult outcomes.[\[19\]](#) Finally, IPW performed quite well in terms of improving the balance in this study; this finding is consistent with prior research.[\[27\]](#) These results highlight the usefulness of implementing several different methods and comparing to ensure results are consistent, as we found in this comparative analysis.

Propensity score methods have been used in previous natural experiment studies of several built environment or policy-based interventions, including park renovations,[\[15\]](#) public transit infrastructure,[\[13\]](#) and a pay-for-performance program for diabetes care.[\[14\]](#) Methods have included weighting[\[15, 33, 34\]](#) and matching.[\[13, 14\]](#) However, these approaches are underutilized. Many previous evaluations of menu labeling policies have used multivariable linear regression models to control for differences between treated and control groups;[\[6–9, 35\]](#) however, regression relies heavily on modeling assumptions regarding the functional form of the relationship between the covariates and the outcome and may be problematic if there are large differences between the two groups.[\[36\]](#) Using IPW, full matching, or other propensity score-based approaches in tandem with multivariable regression may be beneficial in such studies for ensuring that the treatment and control groups are as similar as possible in order to strengthen causal claims.

This study had several limitations. First, the study included a convenience sample of treatment and control restaurant outlets in both urban and suburban areas; however, all data were collected from the same restaurant chain in one part of the U.S. during a single month, which limited variation in potential restaurant-level confounders such as price, menu offerings, and restaurant marketing and promotions. Additionally, the propensity score-based analyses relied on the assumption that there were no unobserved confounding variables, which could be more fully evaluated using a sensitivity analysis. However, we included a comprehensive list of potential confounders in the analyses. Propensity score matching on observed covariates also matches on unobserved covariates that are correlated with observed covariates, but would not account for imbalance in other unobserved covariates.

Conclusion

In this menu labeling natural experiment, full matching on propensity scores and inverse probability weighting were useful methods for improving balance between study groups without a reduction in sample size. The association between menu labeling and purchasing restaurant items lower in calories, fat, carbohydrates, and sodium was slightly attenuated when compared to results that did not use propensity score methods, but results were substantively similar.

Author Contributions

Conceived and designed the experiments: SLM BKL AHA. Performed the experiments: SLM. Analyzed the data: SLM. Contributed reagents/materials/analysis tools: AHA BKL. Wrote the paper: SLM AHA BKL.

References

1. Swinburn BA, Sacks G, Hall KD, McPherson K, Finegood DT, Moodie ML, Gortmaker SL: The global obesity pandemic: shaped by global drivers and local environments. *Lancet* 2011, 378(9793):804–814. doi: [10.1016/S0140-6736\(11\)60813-1](https://doi.org/10.1016/S0140-6736(11)60813-1) PMID: [21872749](https://pubmed.ncbi.nlm.nih.gov/21872749/)
2. Bleich S, Cutler D, Murray C, Adams A: Why is the developed world obese? *Annu Rev Public Health* 2008, 29:273–295. doi: [10.1146/annurev.publhealth.29.020907.090954](https://doi.org/10.1146/annurev.publhealth.29.020907.090954) PMID: [18173389](https://pubmed.ncbi.nlm.nih.gov/18173389/)
3. Guthrie JF, Lin BH, Frazao E: Role of food prepared away from home in the American diet, 1977–78 versus 1994–96: changes and consequences. *Journal of nutrition education and behavior* 2002, 34(3):140–150. PMID: [12047838](https://pubmed.ncbi.nlm.nih.gov/12047838/)
4. Elbel B, Gyamfi J, Kersh R: Child and adolescent fast-food choice and the influence of calorie labeling: a natural experiment. *International journal of obesity* 2011, 35(4):493–500. doi: [10.1038/ijo.2011.4](https://doi.org/10.1038/ijo.2011.4) PMID: [21326209](https://pubmed.ncbi.nlm.nih.gov/21326209/)
5. Auchincloss AH, Mallya GG, Leonberg BL, Ricchezza A, Glanz K, Schwarz DF: Customer responses to mandatory menu labeling at full-service restaurants. *American journal of preventive medicine* 2013, 45(6):710–719. doi: [10.1016/j.amepre.2013.07.014](https://doi.org/10.1016/j.amepre.2013.07.014) PMID: [24237912](https://pubmed.ncbi.nlm.nih.gov/24237912/)
6. Vadiveloo MK, Dixon LB, Elbel B: Consumer purchasing patterns in response to calorie labeling legislation in New York City. *The international journal of behavioral nutrition and physical activity* 2011, 8:51. doi: [10.1186/1479-5868-8-51](https://doi.org/10.1186/1479-5868-8-51) PMID: [21619632](https://pubmed.ncbi.nlm.nih.gov/21619632/)
7. Tandon PS, Zhou C, Chan NL, Lozano P, Couch SC, Glanz K, et al. The impact of menu labeling on fast-food purchases for children and parents. *American journal of preventive medicine* 2011, 41(4):434–438. doi: [10.1016/j.amepre.2011.06.033](https://doi.org/10.1016/j.amepre.2011.06.033) PMID: [21961472](https://pubmed.ncbi.nlm.nih.gov/21961472/)
8. Finkelstein EA, Strombotne KL, Chan NL, Krieger J: Mandatory menu labeling in one fast-food chain in King County, Washington. *American journal of preventive medicine* 2011, 40(2):122–127. doi: [10.1016/j.amepre.2010.10.019](https://doi.org/10.1016/j.amepre.2010.10.019) PMID: [21238859](https://pubmed.ncbi.nlm.nih.gov/21238859/)
9. Dumanovsky T, Huang CY, Nonas CA, Matte TD, Bassett MT, Silver LD: Changes in energy content of lunchtime purchases from fast food restaurants after introduction of calorie labelling: cross sectional customer surveys. *Bmj* 2011, 343:d4464. doi: [10.1136/bmj.d4464](https://doi.org/10.1136/bmj.d4464) PMID: [21791497](https://pubmed.ncbi.nlm.nih.gov/21791497/)
10. PDPH: Philadelphia Menu Labeling Ordinance [Code §§ 6–102, 6–308] <http://www.phila.gov/health/pdfs/MenuLabelinguideFINAL2010-27.pdf>. In. Philadelphia: Philadelphia Department of Public Health; 2010.

11. Rosenbaum PR, Rubin DB: The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika* 1983, 70(1):41–55.
12. Conniffe D, Gash V, O'Connell PJ: Evaluating state programmes: "Natural experiments" and Propensity Scores. *Econ Soc Rev* 2000, 31(4):283–308.
13. Cerda M, Morenoff JD, Hansen BB, Tessari Hicks KJ, Duque LF, Restrepo A, et al. Reducing violence by transforming neighborhoods: a natural experiment in Medellin, Colombia. *American journal of epidemiology* 2012, 175(10):1045–1053. doi: [10.1093/aje/kwr428](https://doi.org/10.1093/aje/kwr428) PMID: [22472117](https://pubmed.ncbi.nlm.nih.gov/22472117/)
14. Cheng SH, Lee TT, Chen CC: A longitudinal examination of a pay-for-performance program for diabetes care: evidence from a natural experiment. *Medical care* 2012, 50(2):109–116. doi: [10.1097/MLR.0b013e31822d5d36](https://doi.org/10.1097/MLR.0b013e31822d5d36) PMID: [22249920](https://pubmed.ncbi.nlm.nih.gov/22249920/)
15. Cohen DA, Marsh T, Williamson S, Golinelli D, McKenzie TL: Impact and cost-effectiveness of family Fitness Zones: a natural experiment in urban public parks. *Health & place* 2012, 18(1):39–45.
16. Austin PC: An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate behavioral research* 2011, 46(3):399–424. PMID: [21818162](https://pubmed.ncbi.nlm.nih.gov/21818162/)
17. Austin PC, Mamdani MM: A comparison of propensity score methods: a case-study estimating the effectiveness of post-AMI statin use. *Stat Med* 2006, 25(12):2084–2106. PMID: [16220490](https://pubmed.ncbi.nlm.nih.gov/16220490/)
18. Hansen BB: Full matching in an observational study of coaching for the SAT. *J Am Stat Assoc* 2004, 99(467):609–618.
19. Stuart EA, Green KM: Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Dev Psychol* 2008, 44(2):395–406. doi: [10.1037/0012-1649.44.2.395](https://doi.org/10.1037/0012-1649.44.2.395) PMID: [18331131](https://pubmed.ncbi.nlm.nih.gov/18331131/)
20. Ho D, Imai K, King G, Stuart EA: Nonparametric preprocessing for parametric causal inference: R Package version 2.2–13; 2007.
21. Hansen BB, Klopfer SO: Optimal full matching and related designs via network flows. *J Comput Graph Stat* 2006, 15(3):609–627.
22. Rosenbaum PR, Rubin DB: Constructing a Control-Group Using Multivariate Matched Sampling Methods That Incorporate the Propensity Score. *Am Stat* 1985, 39(1):33–38.
23. Harder VS, Stuart EA, Anthony JC: Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychological methods* 2010, 15(3):234–249. doi: [10.1037/a0019623](https://doi.org/10.1037/a0019623) PMID: [20822250](https://pubmed.ncbi.nlm.nih.gov/20822250/)
24. Austin PC: Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies. *Pharm Stat* 2011, 10(2):150–161. doi: [10.1002/pst.433](https://doi.org/10.1002/pst.433) PMID: [20925139](https://pubmed.ncbi.nlm.nih.gov/20925139/)
25. Rosenbaum PR: A Characterization of Optimal Designs for Observational Studies. *J Roy Stat Soc B Met* 1991, 53(3):597–610.
26. Stuart EA: Matching Methods for Causal Inference: A Review and a Look Forward. *Stat Sci* 2010, 25(1):1–21. PMID: [20871802](https://pubmed.ncbi.nlm.nih.gov/20871802/)
27. Austin PC: The Relative Ability of Different Propensity Score Methods to Balance Measured Covariates Between Treated and Untreated Subjects in Observational Studies. *Med Decis Making* 2009, 29(6):661–677. doi: [10.1177/0272989X09341755](https://doi.org/10.1177/0272989X09341755) PMID: [19684288](https://pubmed.ncbi.nlm.nih.gov/19684288/)
28. Rosenbaum PR: Model-based direct adjustment. *The Journal of the American Statistical Association* 1987, 82(398):387–394.
29. Funk MJ, Westreich D, Wiesen C, Sturmer T, Brookhart MA, Davidian M: Doubly robust estimation of causal effects. *American journal of epidemiology* 2011, 173(7):761–767. doi: [10.1093/aje/kwq439](https://doi.org/10.1093/aje/kwq439) PMID: [21385832](https://pubmed.ncbi.nlm.nih.gov/21385832/)
30. R Development Core Team: R: A Language and Environment for Statistical Computing; 2008.
31. Rosenbaum PR, Rubin DB: The Bias Due to Incomplete Matching. *Biometrics* 1985, 41(1):103–116. PMID: [4005368](https://pubmed.ncbi.nlm.nih.gov/4005368/)
32. Gu SS, Rosenbaum PR: Comparison of Multivariate Matching Methods: Structures, Distances, and Algorithms. *J Comput Graph Stat* 1993, 2(4):405–420.
33. Elbel B, Mijanovich T, Dixon LB, Abrams C, Weitzman B, Kersh R, et al. Calorie labeling, fast food purchasing and restaurant visits. *Obesity* 2013, 21(11):2172–2179. doi: [10.1002/oby.20550](https://doi.org/10.1002/oby.20550) PMID: [24136905](https://pubmed.ncbi.nlm.nih.gov/24136905/)
34. MacDonald JM, Stokes RJ, Cohen DA, Kofner A, Ridgeway GK: The Effect of Light Rail Transit on Body Mass Index and Physical Activity. *American journal of preventive medicine* 2010, 39(2):105–112. doi: [10.1016/j.amepre.2010.03.016](https://doi.org/10.1016/j.amepre.2010.03.016) PMID: [20621257](https://pubmed.ncbi.nlm.nih.gov/20621257/)

35. Elbel B, Kersh R, Brescoll VL, Dixon LB: Calorie labeling and food choices: a first look at the effects on low-income people in New York City. *Health affairs* 2009, 28(6):w1110–1121. doi: [10.1377/hlthaff.28.6.w1110](https://doi.org/10.1377/hlthaff.28.6.w1110) PMID: [19808705](https://pubmed.ncbi.nlm.nih.gov/19808705/)
36. Rubin DB: Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997, 127(8 Pt 2):757–763. PMID: [9382394](https://pubmed.ncbi.nlm.nih.gov/9382394/)