See corresponding editorial on page 1124.

Machine learning as a strategy to account for dietary synergy: an illustration based on dietary intake and adverse pregnancy outcomes

Lisa M Bodnar,[1,2,3] Abigail R Cartus,[1] Sharon I Kirkpatrick,[4] Katherine P Himes,[2,3] Edward H Kennedy,[5] Hyagriv N Simhan,[2,3] William A Grobman,[6] Jennifer Y Duffy,[7] Robert M Silver,[8] Samuel Parry,[9] and Ashley I Naimi[1]

[1]Department of Epidemiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, USA; [2]Department of Obstetrics, Gynecology, and Reproductive Sciences, School of Medicine, University of Pittsburgh, Pittsburgh, PA, USA; [3]Magee-Womens Research Institute, Pittsburgh, PA, USA; [4]School of Public Health and Health Systems, University of Waterloo, Waterloo, Ontario, Canada; [5]Department of Statistics & Data Science, Carnegie Mellon University, Pittsburgh, PA, USA; [6]Department of Obstetrics and Gynecology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA; [7]Department of Obstetrics & Gynecology, School of Medicine, University of California, Irvine, Irvine, CA, USA; [8]Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, UT, USA; and [9]Department of Obstetrics and Gynecology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

ABSTRACT

Background: Conventional analytic approaches for studying diet patterns assume no dietary synergy, which can lead to bias if incorrectly modeled. Machine learning algorithms can overcome these limitations.

Objectives: We estimated associations between fruit and vegetable intake relative to total energy intake and adverse pregnancy outcomes using targeted maximum likelihood estimation (TMLE) paired with the ensemble machine learning algorithm Super Learner, and compared these with results generated from multivariable logistic regression.

Methods: We used data from 7572 women in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be. Usual daily periconceptional intake of total fruits and total vegetables was estimated from an FFQ. We calculated the marginal risk of preterm birth, small-for-gestational-age (SGA) birth, gestational diabetes, and pre-eclampsia according to density of fruits and vegetables (cups/1000 kcal) ≥80th percentile compared with <80th percentile using multivariable logistic regression and Super Learner with TMLE. Models were adjusted for confounders, including other Healthy Eating Index-2010 components.

Results: Using logistic regression, higher fruit and high vegetable densities were associated with 1.1% and 1.4% reductions in pre-eclampsia risk compared with lower densities, respectively. They were not associated with the 3 other outcomes. Using Super Learner with TMLE, high fruit and vegetable densities were associated with fewer cases of preterm birth (–4.0; 95% CI: −4.9, −3.0 and −3.7; 95% CI: −5.0, −2.3), SGA (−1.7; 95% CI: −2.9, −0.51 and −3.8; 95% CI: −5.0, −2.5), and pre-eclampsia (−3.2; 95% CI: −4.2, −2.2 and −4.0; 95% CI: −5.2, −2.7) per 100 births, respectively, and high vegetable densities were associated with a 0.9% increase in risk of gestational diabetes.

Conclusions: The differences in results between Super Learner with TMLE and logistic regression suggest that dietary synergy, which is accounted for in machine learning, may play a role in pregnancy outcomes. This innovative methodology for analyzing dietary data has the potential to advance the study of diet patterns. *Am J Clin Nutr* 2020;111:1235–1243.

Supported by *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD) grants U10 HD063036 (to RTI International), U10 HD063072 (to Case Western Reserve University), U10 HD063047 (to Columbia University), U10 HD063037 (to Indiana University), U10 HD063041 (to University of Pittsburgh), U10 HD063020 (to Northwestern University), U10 HD063046 (to University of California Irvine and WG), U10 HD063048 (to University of Pennsylvania), and U10 HD063053 (to University of Utah).

The study sponsor had no role in the study design; the collection, analysis, and interpretation of data; writing the report; or the decision to submit the report for publication.

Supplemental Figure 1, Supplemental Material, and Supplemental Tables 1 and 2 are available from the "Supplementary data" link in the online posting of the article and from the same link in the online table of contents at https://academic.oup.com/ajcn/.

Data described in the article and code book will not be made available until it is released to the public by the *Eunice Kennedy Shriver* NICHD. The analytic code is provided in the online material.

Address correspondence to LMB (e-mail: lbodnar@pitt.edu).

Abbreviations used: GTT, glucose tolerance testing; nuMoM2b, Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be; RD, risk difference; SGA, small-for-gestational-age; TMLE, targeted maximum likelihood estimation.

Received October 9, 2019. Accepted for publication January 31, 2020.

First published online February 28, 2020; doi: https://doi.org/10.1093/ajcn/nqaa027.

 Copyright © The Author(s) on behalf of the American Society for Nutrition 2020.

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

Keywords: dietary patterns, pregnant women, pregnancy, birth, machine learning, synergy

Introduction

In 2016, over half a million US deaths were attributable to poor dietary patterns (1), emphasizing the need for science-based nutrition recommendations to optimize health. The Dietary Guidelines for Americans, which provides dietary advice for health promotion and disease prevention (2), has received substantial criticism (3–6). Some of the criticism stems from the relative scarcity of evidence available from whole-diet interventions to inform the recommendations (7). Whole-diet interventions, such as the Dietary Approaches to Stop Hypertension trial (8, 9), rigorously test a dietary pattern in its totality, which is the conceptually relevant exposure of interest (10). In the absence of trials, the guidelines rely on the larger body of observational studies examining disease risk associated with dietary patterns captured using self-reported intake.

However, evaluating a complex, multidimensional exposure such as dietary patterns, which constitute consumption of an array of foods and beverages in different amounts and combinations, is difficult. One major challenge is how to account for synergy in dietary patterns (11). Laboratory research shows that eating certain foods in combination has synergistic health effects. For example, a combination of 5 different berries has antioxidant effects greater than the sum of the effects of the individual berries (12). Similar synergy occurs with a combination of broccoli and tomatoes on tumor growth (13). Synergy may be particularly likely when foods are combined across food groups (e.g., legumes and fruit) (14). Antagonistic effects have also been observed (14). Despite these elegant laboratory data, nutritional epidemiology lags behind in accounting for synergy.

The most commonly used methods for the analyses of self-reported dietary data implicitly assume no synergy. Dietary variables are often constructed as diet index scores, factor scores (from factor or principal component analysis), or clusters (from cluster analysis) (15–17). Although these constructs attempt to account for multidimensionality with a focus on dietary patterns rather than specific foods or nutrients, they rely entirely on investigator background knowledge to correctly identify relevant interactions among dietary components a priori. In other words, investigators must manually code all relevant interactions between dietary variables or between a dietary variable and a covariate before running the model (18–20). Yet, knowledge of such interactions is almost never available. Furthermore, ignoring or misspecifying these interactions can lead to biased effect estimates (18–20). Parametric regression also imposes strict assumptions about the nature of variables' relations to one another. These assumptions lead to bias if incorrectly specified, and yet are not typically subject to further investigation.

Machine learning methods can overcome these limitations. They can more optimally handle synergy among dietary components using automated data-adaptive strategies that discover key interactions among variables (19, 21, 22). Machine learning algorithms can also better account for a number of other specifications and assumptions by more flexibly modeling the interrelation among dietary components (23). However, machine learning has only recently been explored in nutritional epidemiology (24–31). Our objective was to demonstrate the application of machine learning in examining associations between dietary intake and pregnancy outcomes—an area with a limited evidence base (32, 33). Specifically, we estimated associations between total fruit and vegetable intake relative to energy intake, accounting for other dietary components considered part of a multidimensional dietary pattern, and 4 adverse pregnancy outcomes [preterm birth, small-for-gestational-age (SGA) birth, gestational diabetes, and pre-eclampsia]. Associations were examined using targeted maximum likelihood estimation (TMLE) paired with an ensemble machine learning algorithm, and compared with results generated from multivariable logistic regression.

Methods

We used data from the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b), a large prospective US pregnancy cohort that has been described in detail previously (34). Briefly, from 2010 to 2013, nuMoM2b enrolled 10,038 women from 8 medical centers across the United States if they had a viable singleton pregnancy, were at 6–13 completed weeks of gestation, and had no previous pregnancy that lasted ≥20 weeks of gestation. At enrollment (6−13 completed weeks of gestation), women completed an FFQ querying usual periconceptional dietary intake. Trained and credentialed study personnel conducted detailed interviews to ascertain data on demographics, medical history, and behaviors, and abstracted data from ultrasound reports conducted by certified sonographers. At least 30 d after delivery, a trained certified chart abstractor recorded final birth outcomes, medical history, and delivery diagnoses and complications. A common protocol and manual of operations was used for all aspects of the study at all sites. Each site's local institutional review board approved the study and all women gave written informed consent. Our analytic sample included 7995 women who delivered at ≥20 weeks of gestation and had complete dietary and birth outcome data (**Supplemental Figure 1**).

Usual dietary intake in the 3 mo before conception was assessed at 6–13 weeks of gestation using a self-administered modified Block 2005 FFQ, which was available in English and Spanish. The instrument assesses 59 nutrients from ∼120 food and beverage items. The FFQ's food list was developed from the NHANES 1999–2002 dietary recall data, and the nutrient database was developed from the USDA Food and Nutrient Database for Dietary Studies (35). Food groups were derived from the MyPyramid Equivalents Database, version 2.0 (36). The questionnaire uses a series of "adjustment" questions to improve the estimation of fat and carbohydrate intake. Portion size is asked for each food, and pictures of portion sizes were given to participants to enhance accuracy. The instrument has been shown to have acceptable validity relative to other self-report assessment tools in many pregnant samples (37–42). The questionnaire was modified to reflect a 3-mo period. Study personnel checked all pages of the FFQ for completeness. Questionnaires were sent to Block Dietary Data Systems (Berkeley, CA) for scanning and nutrient analysis using software developed at the National Cancer Institute (43).

The 2 main exposures of interest in our analysis were total fruits and total vegetables, defined as densities as per the

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

construction of the Healthy Eating Index-2010 (44). Healthy Eating Index-2010 scores and components were calculated by Block Dietary Data Systems using files that disaggregated foods into their component parts. Specifically, usual daily intakes of fruits and vegetables were expressed relative to energy as cups per 1000 kcal. Other Healthy Eating Index-2010 components, including whole grains, dairy products, total protein foods, seafood and plant proteins, fatty acids, refined grains, sodium, and "empty" calories, were considered as part of a multidimensional dietary pattern, and were included in the models as confounders (44). Percentage of empty calories was calculated by summing the energy provided by added sugars, solid fats, and excess alcohol intake (alcohol intake > 13 g/1000 kcal) and dividing by the total daily energy intake (44).

Gestational age was determined by applying the algorithm defined by the nuMoM2b investigators (34). Preterm birth was defined as delivery of a liveborn or stillborn infant between 20 + 0 and 36 + 6 weeks of gestation. Newborns were classified as being SGA if their birth weight was <10th percentile for gestational age at delivery (45, 46).

Gestational diabetes was defined as 1 of the following glucose tolerance testing (GTT) criteria: *1*) 3-h 100-g GTT with 2 values from the following: fasting ≥95 mg/dL, 1-h ≥180 mg/dL, 2-h ≥155 mg/dL, or 3-h ≥140 mg/dL; *2*) 2-h 75-g GTT with 1 value from the following: fasting ≥92 mg/dL, 1-h ≥180 mg/dL, or 2-h ≥153 mg/dL; or *3*) 50-g GTT with a 1-h value ≥200 mg/dL if no fasting 3-h or 2-h GTT was performed (47). GTT was performed as part of routine clinical care.

Detailed definitions of pregnancy hypertensive disorders used in the nuMoM2b cohort have been published (47). Briefly, preeclampsia was the following symptoms occurring at ≥20 weeks of gestation through 14 d: gestational hypertension (≥140 mm Hg systolic or ≥90 mm Hg diastolic blood pressure on 2 occasions ≥6 h apart or 1 occasion with subsequent antihypertensive therapy, excluding blood pressures recorded during the second stage of labor) and proteinuria (≥300 mg/24-h collection or protein:creatinine ratio ≥ 0.3 or dipstick ≥ 2), thrombocytopenia (platelet count < 100,000/mm^3), or pulmonary edema. Preeclampsia included superimposed pre-eclampsia or eclampsia, regardless of the timing of onset. Cases that presented atypically and were difficult to classify according to study criteria were adjudicated by review of clinical data by the principal investigators and final classification was reached by their consensus judgment.

At enrollment, women self-reported their highest level of education, which was categorized as high school or less (less than high school, or high school graduate or equivalent), some college (some college credit, but no degree, or associate/technical degree), college graduate (bachelor's degree), or graduate degree (master's, doctorate, or professional degree). Self-reported race/ethnicity was classified as non-Hispanic white, non-Hispanic black, Hispanic, or other. Other information self-reported at the first visit included marital status, smoking before pregnancy, and medical insurance. At the initial visit, women had their weight measured using an electronic or balance scale while wearing only light clothes and no shoes, and height measured using a stadiometer or measuring tape. Early pregnancy BMI was calculated as self-reported weight divided by measured height squared (kg/m^2). We classified women as underweight (<18.5), normal weight (18.5−24.9), or affected by overweight (25−29.9) or obesity (≥30) (48).

Statistical analysis

We contrasted the marginally adjusted risks of each adverse outcome that would be observed under 2 different dietary exposure scenarios. For example, what would the risk be if all women consumed vegetables at ≥80th percentile of total vegetable consumption, compared with <80th percentile? Our exposures were total fruits and total vegetables, relative to energy intake. We categorized total fruits and total vegetables as ≥80th percentile or <80th percentile of the sample's distribution of each food group. We chose this cutoff because it corresponds to the highest quintile, which is often the primary exposure in nutritional epidemiology. Each exposure association was adjusted for several other Healthy Eating Index-2010 components conceptualized as part of a multidimensional dietary pattern, as aforementioned. Furthermore, all contrasts were adjusted for a set of confounders identified via directed acyclic graphs (49, 50), including maternal race/ethnicity, age, smoking, education, prepregnancy BMI, marital status, and insurance. We performed a sensitivity analysis by limiting the sample to women whose energy intake was from the 5th to the 95th percentiles of the distribution.

We estimated associations of fruit and vegetable density with adverse pregnancy outcomes using multivariable logistic regression and machine learning with automated data-adaptive strategies (Super Learner) via TMLE. Logistic regression requires correctly identifying and coding all relevant interactions between the exposure of interest and all other covariates included in the model (18–20). We did not include interaction terms because we had no prior knowledge of dietary synergy on pregnancy outcomes. For both approaches, we quantified exposure effects by calculating the absolute differences in risk. We multiplied risk differences (RDs) and their corresponding CIs by 100 to estimate the number of excess (or prevented) cases of the adverse outcome.

We included in our ensemble learner a prespecified library of algorithms with tuning parameters. This library included *1*) random forests with minimum node size 500 and 2500 trees, and 2, 3, and 4 predictor variables selected at random for each split (ranger), and sampling with and without replacement; *2*) extreme gradient boosting (xgboost) with maximum tree depth of 4, 5, or 6 and shrinkage parameters of 0.01, 0.001, or 0.0001; *3*) Lasso and elastic-net regularized generalized linear models (glmnet) with elastic net mixing parameter $\alpha = 0.0$ (ridge penalty), 0.2, 0.4, 0.6, 0.8, or 1.0 (Lasso penalty); *4*) k-nearest-neighbors with 5, 10, and 50 nearest neighbors (kernelKNN); *5*) classification and regression trees with default tuning parameters (rpart); *6*) generalized linear models (glm); and *7*) simple mean. To reduce the potential for overfitting, each ensemble learner was fit using 10-fold cross-validation. We calculated the weights and the cross-validated mean squared errors for all algorithms in each Super Learner. We chose 10-fold cross-validation because increasing the number of folds would have led to excessive computing time.

Any machine learning algorithm used to quantify exposure effects may be subject to problems induced by the "curse of dimensionality" (51, 52). These problems include biased effect estimation and poor CI coverage. To address these problems, we implemented our machine learning algorithms using TMLE, a doubly robust, maximum likelihood–based method for parameter

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

TABLE 1 Characteristics of 7572 deliveries in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be

	n (%)
Maternal age, y	
<25	2443 (32)
25–34	4394 (58)
≥35	735 (10)
Maternal race/ethnicity	
Non-Hispanic white	4852 (64)
Non-Hispanic black	821 (11)
Hispanic	1227 (16)
Other	672 (9)
Maternal education	
High school or less	1313 (17)
Some college	1358 (18)
College graduate	3006 (40)
Graduate degree	1895 (25)
Prepregnancy BMI	
Underweight	296 (4)
Normal weight	4280 (56)
Overweight	1638 (22)
Obese	1358 (18)
Smoking status	
Nonsmoker	6322 (83)
Smoker	1250 (17)
Marital status	
Not married	2713 (36)
Married	4859 (64)
Insurance at delivery	
Private	5748 (76)
Public	1824 (24)

estimation. This doubly robust property ensures that parameter estimates generated by TMLE will be unbiased if ≥1 of the exposure or outcome mechanisms is consistently estimated (53). We and others have previously shown that doubly robust estimators like TMLE are less susceptible to the problems that result from the curse of dimensionality (52, 54, 55).

All analyses were conducted with R this version 3.6.1 and Stata version 14. The R code for the logistic regression and Super Learner with TMLE is provided in the **Supplemental Material**.

Results

Most women in the analytic sample were 25–34 y old, non-Hispanic white, college-educated, normal weight, nonsmokers, married, and had private health insurance (**Table 1**). Preterm birth, pre-eclampsia, gestational diabetes, and SGA birth occurred in 8.0%, 8.4%, 4.8%, and 11% of the cohort, respectively.

Approximately 7% of women reported usual intake ≥80th percentile of both fruits (≥1.2 cups/1000 kcal) and vegetables (≥1.3 cups/1000 kcal), whereas 13% had ≥80th-percentile intakes of fruit only and 13% had ≥80th-percentile intakes of vegetables only. Women with high intakes of either fruits or vegetables relative to energy intake were more likely than their counterparts to be older, college educated, normal weight, nonsmokers, married, non-Hispanic white (among those with high vegetables only), and to have private health insurance (**Table 2**). Those with usual intakes ≥80th percentile for fruit had higher mean intakes of vegetables, seafood and plant proteins, and beneficial fatty acids, and lower mean intakes of refined grains and empty calories than women who consumed less fruit. Women with intakes ≥80th percentile of vegetables had higher mean intakes of fruit, seafood and plant proteins, beneficial fatty acids, and sodium, and lower intakes of refined grains and empty calories than those who reported lower vegetable intake. There were no important differences in intake of dairy, total protein foods, or whole grains by categories of fruit or vegetable intake.

The unadjusted incidence of each adverse pregnancy outcome was higher among women who consumed <80th percentile of total fruits and total vegetables than among women consuming ≥80th percentile (**Table 3**).

After adjustment for maternal age, race/ethnicity, education, marital status, insurance, prepregnancy BMI, smoking, and dietary components, associations between the density of total fruits and total vegetables within the diet and preterm birth, SGA birth, and gestational diabetes were null when modeled using logistic regression (**Figure 1**, **Table 4**). In contrast, Super Learner with TMLE produced stronger and more precise effect estimates for preterm birth and SGA. For instance, compared with women who reported consuming <80th percentile of total fruit, women reporting dietary patterns with higher fruit density had 4 fewer preterm births for every 100 women in the sample (RD: −4.0; 95% CI: −4.9, −3.0) using Super Learner with TMLE. The results from logistic regression differed by nearly an order of magnitude (RD: −0.67; 95% CI: −2.4, 1.0). Pre-eclampsia risk was lower among women who reported high fruit and high vegetable intakes than among their counterparts based on both modeling techniques. However, the estimates from Super Learner with TMLE were stronger and more precise (Figure 1).

For all Super Learners, there was no single algorithm that provided the best fit of the data (**Supplemental Tables 1** and **2**). This highlights the importance of an ensemble machine learning approach. For example, for the Super Learner estimation of the outcome mechanism for the relation between fruit density and preterm birth, generalized linear models accounted for ∼41% of the fit and extreme gradient boosting accounted for 33%, whereas the remainder was accounted for by random forests and k-nearest neighbors.

These results did not meaningfully differ when we limited the sample to women with usual energy intake from the 5th to the 95th percentiles of the distribution (data not shown).

Discussion

Although policy makers recognized the importance of dietary synergy when developing the Dietary Guidelines for Americans (2), the vast majority of nutritional epidemiology has not accounted for synergy. Thus, it is not surprising that our results relating fruit and vegetable density with 4 adverse pregnancy outcomes led to markedly different conclusions depending on the analytic approach. We observed predominantly null associations generated from logistic regression models, whereas Super Learner with TMLE produced effect estimates with less variation that suggested protective associations for diets high in fruits and vegetables relative to energy on risk of preterm birth, SGA birth, and pre-eclampsia. The higher risk of gestational diabetes with high vegetable intake could be driven by certain subgroups of vegetables (e.g., potatoes and French fries) and warrants further exploration. Although weak or null

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

TABLE 2 Characteristics of 7572 deliveries in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be according to usual daily fruit and vegetable densities relative to energy intake in the periconceptional period[1]

	Total fruit density		Total vegetable density	
	<80th percentile (<1.2 cups/1000 kcal) ($n = 6058$)	≥80th percentile (≥1.2 cups/1000 kcal) ($n = 1514$)	<80th percentile (<1.2 cups/1000 kcal) ($n = 6058$)	≥80th percentile (≥1.2 cups/1000 kcal) ($n = 1514$)
Maternal age, y				
<25	34	26	37	13
25–34	57	63	55	72
≥35	9	11	8	15
Maternal race/ethnicity				
Non-Hispanic white	64	65	61	77
Non-Hispanic black	12	9	13	3
Hispanic	16	18	18	11
Other	9	9	9	9
Maternal education				
High school or less	19	11	20	6
Some college	19	15	20	10
College graduate	39	43	39	43
Graduate degree	24	30	21	41
Prepregnancy BMI				
Underweight	4	4	4	3
Normal weight	55	62	55	62
Overweight	22	19	22	22
Obese	19	15	19	13
Smoking status				
Nonsmoker	82	91	82	89
Smoker	18	9	18	11
Marital status				
Not married	38	27	41	17
Married	62	73	59	83
Insurance at delivery				
Private	75	81	72	90
Public	25	19	28	10
Total fruit, cups/1000 kcal	0.63 ± 0.30	1.7 ± 0.44	0.77 ± 0.50	1.1 ± 0.58
Total vegetables, cups/1000 kcal	0.88 ± 0.48	1.2 ± 0.63	0.73 ± 0.27	1.8 ± 0.51
Dairy, cups/1000 kcal	0.88 ± 0.47	0.88 ± 0.50	0.89 ± 0.49	0.82 ± 0.41
Total protein foods, oz/1000 kcal	2.4 ± 0.76	2.2 ± 0.80	2.4 ± 0.74	2.5 ± 0.85
Seafood and plant proteins, oz/1000 kcal	0.83 ± 0.57	0.97 ± 0.64	0.79 ± 0.54	1.2 ± 0.70
Fatty acids[2]	1.4 ± 1.1	1.7 ± 1.13	1.4 ± 1.2	1.7 ± 0.90
Refined grains, oz/1000 kcal	2.3 ± 0.73	1.9 ± 0.64	2.3 ± 0.73	1.9 ± 0.67
Whole grains, oz/1000 kcal	0.62 ± 0.44	0.68 ± 0.44	0.62 ± 0.44	0.68 ± 0.43
Sodium, g/1000 kcal	1.6 ± 0.26	1.5 ± 0.26	1.6 ± 0.24	1.8 ± 0.25
Empty calories, % of energy	33 ± 8.6	27 ± 6.7	33 ± 8.4	26 ± 6.1

[1]Values are means ± SDs or percentages. One cup is equivalent to 237 mL; 1 oz is equivalent to 30 mL.
[2]Ratio of PUFAs and MUFAs to SFAs.

associations in nutritional epidemiology are often attributed to measurement errors (56), other factors, including misspecifying the underlying relations among dietary components and how each plays a role in affecting pregnancy outcomes, likely also play a role.

We are aware of only 1 other nutrition study to compare parametric approaches with machine learning. Researchers compared linear regression with k-nearest neighbor algorithm and random-forest decision tree in their ability to accurately classify cardiometabolic risk based on dietary data (27). They found that machine learning approaches correctly classified 38% of individuals compared with 6% with linear regression. We previously compared the relative performance of Bayesian additive regression trees, generalized boosted models, or random forests to parametric regression models and demonstrated a much higher predictive accuracy with machine learning methods than with regression, as well as important differences in the strength of associations (57).

Machine learning has become popular in many scientific fields plagued by problems with high-dimensional data. It is only beginning to be explored in nutrition, but several elegant studies demonstrate the tremendous potential of this approach (24–31). For example, investigators used a machine learning algorithm to combine data on dietary intake, biomarkers, anthropometrics, physical activity, and gut microbiota to accurately predict personalized postprandial glycemic response (29).

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

TABLE 3 Incidence of adverse pregnancy outcomes by dietary variables in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be according to usual intake of fruit and vegetables in the periconceptional period[1]

Dietary exposure	Population at risk, *n*	Preterm birth cases, *n* (%)	Small-for-gestational-age birth cases, *n* (%)	Gestational diabetes cases, *n* (%)	Pre-eclampsia cases, *n* (%)
Total fruit density					
<80th percentile	6058	468 (7.7)	702 (11.6)	295 (4.9)	530 (8.8)
≥80th percentile	1514	93 (6.1)	165 (10.9)	67 (4.4)	106 (7.0)
Total vegetable density					
<80th percentile	6058	460 (7.7)	708 (11.8)	292 (4.9)	539 (9.0)
≥80th percentile	1514	101 (6.4)	159 (10.1)	70 (4.4)	97 (6.2)

[1]The 80th percentiles are 1.2 cups fruit/1000 kcal and 1.3 cups vegetables/1000 kcal. One cup is equivalent to 237 mL.

Personalized nutrition (customized nutrition advice to optimize individual health) was also explored in a study integrating the Healthy Eating Index, folate metabolic genes, and other risk factors in an ensemble machine learning procedure to predict colorectal cancer (30). The same research team used machine learning to evaluate predictors of the Healthy Eating Index and glycemic index (31).

Although important and groundbreaking, previous machine learning work in nutritional epidemiology has focused on risk prediction or classification (i.e., using a set of variables to predict the outcome well). Our research extends the field to demonstrate an approach for effect estimation (i.e., estimating the effects of an exposure on the distribution of disease, while controlling for other variables), which is of the greatest interest in developing national dietary guidance. Machine learning methods may be subject to biased effect estimation and poor CI coverage (51, 52). However, our implementation of Super Learner using TMLE avoids this problem with doubly robust techniques,

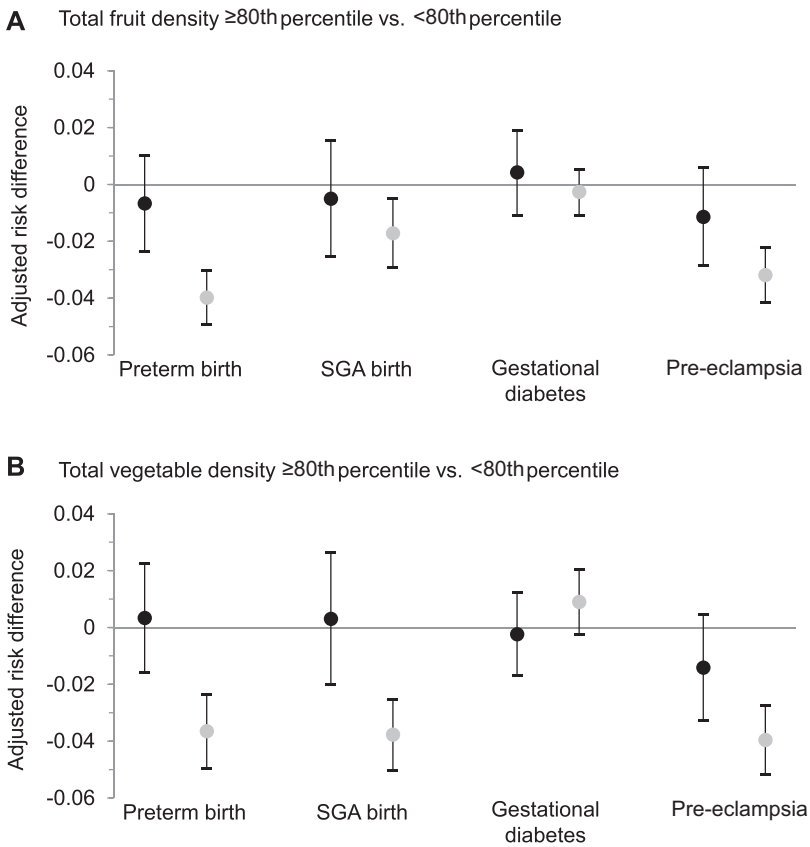

FIGURE 1 Associations between fruit and vegetable density and risk of adverse pregnancy outcomes in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be ($n = 7252$). (A) Adjusted differences in risk (95% CIs) of preterm birth, SGA birth, gestational diabetes, and pre-eclampsia between total fruit intake relative to energy ≥80th percentile (≥1.2 cups/1000 kcal) ($n = 1514$) compared with <80th percentile ($n = 6058$) and (B) total vegetable intake relative to energy ≥80th percentile (≥1.3 cups/1000 kcal) ($n = 1514$) compared with <80th percentile ($n = 6058$). Point estimates in black were generated from multivariable logistic regression. Point estimates in gray were generated from Super Learner with targeted maximum likelihood estimation. All models were adjusted for maternal age, race/ethnicity, education, marital status, smoking status, prepregnancy BMI, insurance, and usual dietary intake of whole grains, dairy products, total protein foods, seafood and plant proteins, fatty acids, refined grains, sodium, and "empty" calories. SGA, small-for-gestational-age.

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

TABLE 4 Association between fruit and vegetable density and risk of adverse pregnancy outcomes in the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be[1]

Dietary exposure	Multivariable logistic regression Adjusted[2] number of excess cases (95% CI)	Super Learner with targeted maximum likelihood estimation Adjusted[2] number of excess cases (95% CI)
Preterm birth		
Total fruit density		
<80th percentile	Referent	Referent
≥80th percentile	−0.67 (−2.4, 1.0)	−4.0 (−4.9, −3.0)
Total vegetable density		
<80th percentile	Referent	Referent
≥80th percentile	0.34 (−1.6, 2.3)	−3.7 (−5.0, −2.3)
Small-for-gestational age birth		
Total fruit density		
<80th percentile	Referent	Referent
≥80th percentile	−0.50 (−2.5, 1.5)	−1.7 (−2.9, −0.51)
Total vegetable density		
<80th percentile	Referent	Referent
≥80th percentile	0.31 (−2.0, 2.6)	−3.8 (−5.0, −2.5)
Gestational diabetes		
Total fruit density		
<80th percentile	Referent	Referent
≥80th percentile	0.42 (−1.1, 1.9)	−0.26 (−1.1, 0.53)
Total vegetable density		
<80th percentile	Referent	Referent
≥80th percentile	−0.23 (−1.7, 1.2)	0.90 (−0.23, 2.0)
Pre-eclampsia		
Total fruit density		
<80th percentile	Referent	Referent
≥80th percentile	−1.1 (−2.9, 0.60)	−3.2 (−4.2, −2.2)
Total vegetable density		
<80th percentile	Referent	Referent
≥80th percentile	−1.4 (−3.3, 0.46)	−4.0 (−5.2, −2.7)

[1] $n = 7252$. The 80th percentiles are 1.2 cups fruit/1000 kcal and 1.3 cups vegetables/1000 kcal. One cup is equivalent to 237 mL.

[2] Adjusted for maternal age, race/ethnicity, education, marital status, smoking status, prepregnancy BMI, insurance, and usual dietary intake of whole grains, dairy products, total protein foods, seafood and plant proteins, fatty acids, refined grains, sodium, and percentage of total calories that are "empty" calories.

allowing us to quantify the targeted associations of interest. A downside to our approach is the reliance on a dichotomous primary exposure (58). We chose to dichotomize total fruits and total vegetables at the 80th percentile of their respective distributions because this corresponds to the cutoff for the upper quintile, often of interest in nutritional epidemiology. Future research should explore other cutoffs as well as the use of continuous measures of dietary intake.

Our work was limited by the use of data from an FFQ, which are affected by systematic measurement error to a greater extent than other self-report methods such as 24-h recalls (59, 60). Our adjustment for energy intake helps to attenuate the potential for bias (61). We restricted our analysis to fruit and vegetable food groups, but of greater interest is a more comprehensive evaluation of foods in the exploration of the global role of dietary patterns on health outcomes. The nuMoM2b study enrolled a racially/ethnically diverse sample from 8 US centers, but we sacrificed some generalizability in using this group because it included only nulliparous women.

Our results show that Super Learner implemented with TMLE can be used with dietary data to generate compelling interpretable results that account for dietary synergy. The complex interactive effects in the diet and the multidimensional nature of dietary information make nutrition data ideally suited for machine learning applications. This innovative methodology for statistical analysis of dietary data has the potential to advance the field of nutritional epidemiology, enhancing the evidence base for recommendations that embrace the whole diet.

We thank Sara Parisi for assistance with data management. The following institutions and researchers compose the Nulliparous Pregnancy Outcomes Study: Monitoring Mothers-to-be (nuMoM2b) Network: Case Western Reserve University/Ohio State University—Brian M Mercer, Jay Iams, Wendy Dalton, Cheryl Latimer, LuAnn Polito; Columbia University/Christiana Care—Matthew K Hoffman, Ronald Wapner, Karin Fuchs, Caroline Torres, Stephanie Lynch, Ameneh Onativia, Michelle DiVito; Indiana University—David M Haas, Tatiana Foroud, Emily Perkins, Shannon Barnes, Alicia Winters, Catherine L McCormick; University of Pittsburgh—Hyagriv N Simhan, Steve N Caritis, Melissa Bickus, Paul D Speer, Stephen P Emery, Ashi R Daftary; Northwestern University—William A Grobman, Alan M Peaceman, Peggy Campbell, Jessica S Shepard, Crystal N Williams; University of California at Irvine—Deborah A Wing, Pathik D Wadhwa, Michael P Nageotte, Pamela J Rumney, Manuel Porto, Valerie Pham; University of Pennsylvania—Samuel Parry, Jack Ludmir, Michal Elovitz, Mary Peters, Brittany Araujo; University of Utah—Robert M Silver, M Sean Esplin, Kelly Vorwaller, Julie Postma, Valerie Morby, Melanie Williams, Linda Meadows; RTI International—Corette B Parker, Matthew A Koch, Deborah W McFadden, Barbara V Alexander, Venkat Yetukuri, Shannon

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

Hunter, Tommy E Holder, Jr, Holly L Franklin, Martha J DeCain, Christopher Griggs; *Eunice Kennedy Shriver* National Institute of Child Health and Human Development—Uma M Reddy, Marian Willinger, Maurice Davis; University of Texas Medical Branch at Galveston—George R Saade.

The authors' responsibilities were as follows—LMB and AIN: designed the research; LMB, ARC, SIK, KPH, and AIN: conducted the research; HNS, WAG, JYD, RMS, and SP: provided essential reagents or provided essential materials and critically reviewed and edited the paper; LMB, ARC, and AIN: analyzed the data or performed statistical analysis; LMB, ARC, SIK, KPH, EHK, and AIN: wrote the paper; LMB: had primary responsibility for the final content; and all authors: read and approved the final manuscript. The authors report no conflicts of interest.

References

1. Mokdad AH, Ballestros K, Echko M, Glenn S, Olsen HE, Mullany E, Lee A, Khan AR, Ahmadi A, Ferrari AJ, et al. The state of US health, 1990–2016: burden of diseases, injuries, and risk factors among US states. JAMA 2018;319(14):1444–72.
2. Dietary Guidelines Advisory Committee. Report of the Dietary Guidelines Advisory Committee on the Dietary Guidelines for Americans, 2010, to the Secretary of Agriculture and the Secretary of Health and Human Services. Washington (DC): USDA and US Department of Health and Human Services; 2010.
3. Nissen SE. U.S. Dietary Guidelines: an evidence-free zone. Ann Intern Med 2016;164(8):558–9.
4. Teicholz N. The scientific report guiding the US dietary guidelines: is it scientific? BMJ 2015;351:h4962.
5. Hite AH, Feinman RD, Guzman GE, Satin M, Schoenfeld PA, Wood RJ. In the face of contradictory evidence: report of the Dietary Guidelines for Americans Committee. Nutrition 2010;26(10):915–24.
6. Marantz PR, Bird ED, Alderman MH. A call for higher standards of evidence for dietary guidelines. Am J Prev Med 2008;34(3):234–40.
7. GBD 2017 Diet Collaborators. Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. Lancet 2019;393(10184):1958–72.
8. de Lorgeril M, Renaud S, Mamelle N, Salen P, Martin JL, Monjaud I, Guidollet J, Touboul P, Delaye J. Mediterranean alpha-linolenic acid-rich diet in secondary prevention of coronary heart disease. Lancet 1994;343(8911):1454–9.
9. Harsha DW, Lin PH, Obarzanek E, Karanja NM, Moore TJ, Caballero B; DASH Collaborative Research Group. Dietary Approaches to Stop Hypertension: a summary of study results. J Am Diet Assoc 1999;99(8 Suppl):S35–9.
10. Jacobs DR Jr, Tapsell LC. Food, not nutrients, is the fundamental unit in nutrition. Nutr Rev 2007;65(10):439–50.
11. Reedy J, Subar AF, George SM, Krebs-Smith SM. Extending methods in dietary patterns research. Nutrients 2018;10(5):571.
12. Zafra-Stone S, Yasmin T, Bagchi M, Chatterjee A, Vinson JA, Bagchi D. Berry anthocyanins as novel antioxidants in human health and disease prevention. Mol Nutr Food Res 2007;51:675–83.
13. Canene-Adams K, Lindshield BL, Wang S, Jeffery EH, Clinton SK, Erdman JW Jr. Combinations of tomato and broccoli enhance antitumor activity in Dunning R3327-H prostate adenocarcinomas. Cancer Res 2007;67(2):836–43.
14. Wang S, Meckling KA, Marcone MF, Kakuda Y, Tsao R. Synergistic, additive, and antagonistic effects of food mixtures on total antioxidant capacities. J Agric Food Chem 2011;59(3):960–8.
15. Schulze MB, Martínez-González MA, Fung TT, Lichtenstein AH, Forouhi NG. Food based dietary patterns and chronic disease prevention. BMJ 2018;361:k2396.
16. Schulze MB, Hoffmann K. Methodological approaches to study dietary patterns in relation to risk of coronary heart disease and stroke. Br J Nutr 2006;95(5):860–9.
17. Krebs-Smith SM, Subar AF, Reedy J. Examining dietary patterns in relation to chronic disease: matching measures and methods to questions of interest. Circulation 2015;132(9):790–3.
18. Yang PY, Yang YH, Zhou BB, Zomaya AY. A review of ensemble methods in bioinformatics. Curr Bioinform 2010;5(4):296–308.
19. Grömping U. Variable importance assessment in regression: linear regression versus random forest. Am Stat 2009;63(4):308–19.
20. García-Magariños M, López-de-Ullibarri I, Cao R, Salas A. Evaluating the ability of tree-based methods and logistic regression for the detection of SNP-SNP interaction. Ann Hum Genet 2009;73(Pt 3):360–9.
21. Hastie T, Tibshirani R, Friedman JJH. The elements of statistical learning. New York: Springer; 2009.
22. Denisko D, Hoffman MM. Classification and interaction in random forests. Proc Natl Acad Sci U S A 2018;115(8):1690–2.
23. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. Eur J Epidemiol 2018;33(5):459–64.
24. Acar E, Gürdeniz G, Khakimov B, Savorani F, Korndal SK, Larsen TM, Engelsen SB, Astrup A, Dragsted LO. Biomarkers of individual foods, and separation of diets using untargeted LC–MS-based plasma metabolomics in a randomized controlled trial. Mol Nutr Food Res 2019;63(1):1800215.
25. Jiang L, Audouze K, Romero Herrera JA, Ängquist LH, Kjærulff SK, Izarzugaza JMG, Tjønneland A, Halkjær J, Overvad K, Sørensen TIA, et al. Conflicting associations between dietary patterns and changes of anthropometric traits across subgroups of middle-aged women and men. Clin Nutr 2020;39(1):265–75.
26. Kanerva N, Kontto J, Erkkola M, Nevalainen J, Männistö S. Suitability of random forest analysis for epidemiological research: exploring sociodemographic and lifestyle-related risk factors of overweight in a cross-sectional design. Scand J Public Health 2018;46(5):557–64.
27. Panaretos D, Koloverou E, Dimopoulos AC, Kouli GM, Vamvakari M, Tzavelas G, Pitsavos C, Panagiotakos DB. A comparison of statistical and machine-learning techniques in evaluating the association between dietary patterns and 10-year cardiometabolic risk (2002–2012): the ATTICA study. Br J Nutr 2018;120(3):326–34.
28. Rosso N, Giabbanelli P. Accurately inferring compliance to five major food guidelines through simplified surveys: applying data mining to the UK National Diet and Nutrition Survey. JMIR Public Health Surveill 2018;4(2):e56.
29. Zeevi D, Korem T, Zmora N, Israeli D, Rothschild D, Weinberger A, Ben-Yacov O, Lador D, Avnit-Sagi T, Lotan-Pompan M, et al. Personalized nutrition by prediction of glycemic responses. Cell 2015;163(5):1079–94.
30. Shiao SPK, Grayson J, Lie A, Yu CH. Personalized nutrition—genes, diet, and related interactive parameters as predictors of cancer in multiethnic colorectal cancer families. Nutrients 2018;10(6):795.
31. Shiao SPK, Grayson J, Lie A, Yu CH. Predictors of the healthy eating index and glycemic index in multi-ethnic colorectal cancer families. Nutrients 2018;10(6):674.
32. Raghavan R, Dreibelbis C, Kingshipp BL, Wong YP, Abrams B, Gernand AD, Rasmussen KM, Siega-Riz AM, Stang J, Casavale KO, et al. Dietary patterns before and during pregnancy and birth outcomes: a systematic review. Am J Clin Nutr 2019;109(Supplement_1):729S–56S.
33. Raghavan R, Dreibelbis C, Kingshipp BL, Wong YP, Abrams B, Gernand AD, Rasmussen KM, Siega-Riz AM, Stang J, Casavale KO, et al. Dietary patterns before and during pregnancy and maternal outcomes: a systematic review. Am J Clin Nutr 2019;109(Supplement 1):705S–28S.
34. Haas DM, Parker CB, Wing DA, Parry S, Grobman WA, Mercer BM, Simhan HN, Hoffman MK, Silver RM, Wadhwa P, et al. A description of the methods of the Nulliparous Pregnancy Outcomes Study: monitoring mothers-to-be (nuMoM2b). Am J Obstet Gynecol 2015;212(4):539.e1–e24.
35. USDA. USDA Food and Nutrient Database for Dietary Studies version 1.0. Beltsville, MD: Agricultural Research Service, Food Surveys Research Group; 2004.
36. Bowman SA, Friday JE, Moshfegh A. MyPyramid Equivalents Database, version 2.0 for USDA Survey Foods, 2003–2004 [Internet]. Beltsville, MD: USDA Agricultural Research Service; 2004 [cited 8 September, 2016]. Available from: http://www.ars.usda.gov/Services/docs.htm?docid=17565.
37. Block G, Hartman AM, Dresser CM, Carroll MD, Gannon J, Gardner L. A data-based approach to diet questionnaire design and testing. Am J Epidemiol 1986;124(3):453–69.
38. Block G, Woods M, Potosky A, Clifford C. Validation of a self-administered diet history questionnaire using multiple diet records. J Clin Epidemiol 1990;43(12):1327–35.

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022

39. Johnson BA, Herring AH, Ibrahim JG, Siega-Riz AM. Structured measurement error in nutritional epidemiology: applications in the Pregnancy, Infection, and Nutrition (PIN) Study. J Am Statist Assoc 2007;102(479):856–66.
40. Mares-Perlman JA, Klein BE, Klein R, Ritter LL, Fisher MR, Freudenheim JL. A diet history questionnaire ranks nutrient intakes in middle-aged and older men and women similarly to multiple food records. J Nutr 1993;123(3):489–501.
41. Boucher B, Cotterchio M, Kreiger N, Nadalin V, Block T, Block G. Validity and reliability of the Block98 food-frequency questionnaire in a sample of Canadian women. Public Health Nutr 2006;9(1):84–93.
42. Block G, Coyle LM, Hartman AM, Scoppa SM. Revision of dietary analysis software for the Health Habits and History Questionnaire. Am J Epidemiol 1994;139(12):1190–6.
43. National Cancer Institute, Epidemiology and Genomics Research Program. Diet∗Calc Analysis Program, version 1.5.0. Bethesda, MD: National Cancer Institute; 2012.
44. Guenther PM, Casavale KO, Reedy J, Kirkpatrick SI, Hiza HA, Kuczynski KJ, Kahle LL, Krebs-Smith SM. Update of the Healthy Eating Index: HEI-2010. J Acad Nutr Diet 2013;113(4):569–80.
45. Lu MS, Chen QZ, He JR, Wei XL, Lu JH, Li SH, Wen XX, Chan FF, Chen NN, Qiu L, et al. Maternal dietary patterns and fetal growth: a large prospective cohort study in China. Nutrients 2016;8(5):257.
46. Alexander GR, Himes JH, Kaufman RB, Mor J, Kogan M. A United States national reference for fetal growth. Obstet Gynecol 1996;87(2):163–8.
47. Facco FL, Parker CB, Reddy UM, Silver RM, Koch MA, Louis JM, Basner RC, Chung JH, Nhan-Chang CL, Pien GW, et al. Association between sleep-disordered breathing and hypertensive disorders of pregnancy and gestational diabetes mellitus. Obstet Gynecol 2017;129(1):31–41.
48. WHO Consultation on Obesity. Obesity: preventing and managing the global epidemic. Geneva, Switzerland: World Health Organization; 2000.
49. Shrier I, Platt RW. Reducing bias through directed acyclic graphs. BMC Med Res Methodol 2008;8:70.
50. Hernán MA, Hernández-Díaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155:176–84.
51. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J. Double/debiased machine learning for treatment and structural parameters. Economet J 2018;21(1):C1–C68.
52. Naimi AI, Kennedy EH. Nonparametric double robustness [Internet]. 2017 [cited 31 October, 2019]. Available from: https://arxivorg/abs/171107137 [statME].
53. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. Am J Epidemiol 2017;185(1):65–73.
54. Díaz I, Carone M, van der Laan MJ. Second-order inference for the mean of a variable missing at random. Int J Biostat 2016;12(1):333–49.
55. Rothe C, Firpo S. Properties of doubly robust estimators when nuisance function are estimated nonparametrically [Internet]. Working paper. 2017 [cited 31 October, 2019]. Available from: https://www.cambridge.org/core/journals/econometric-theory/article/properties-of-doubly-robust-estimators-when-nuisance-functions-are-estimated-nonparametrically/A9BA1449CD982BC35C245BFEE680759F.
56. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 2011;103(14):1086–92.
57. Naimi AI, Platt RW, Larkin JC. Machine learning for fetal growth prediction. Epidemiology 2018;29(2):290–8.
58. Gruber S, van der Laan M. tmle: an R package for targeted maximum likelihood estimation. J Stat Soft 2012;51(13):35.
59. Kipnis V, Subar AF, Midthune D, Freedman LS, Ballard-Barbash R, Troiano RP, Bingham S, Schoeller DA, Schatzkin A, Carroll RJ. Structure of dietary measurement error: results of the OPEN biomarker study. Am J Epidemiol 2003;158(1):14–21; discussion 22–6.
60. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN study. Am J Epidemiol 2003;158(1):1–13.
61. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, Thompson FE, Potischman N, Guenther PM, Tarasuk V, et al. Addressing current criticism regarding the value of self-report dietary data. J Nutr 2015;145(12):2639–45.

Downloaded from https://academic.oup.com/ajcn/article/111/6/1235/5763095 by guest on 30 April 2022