The Journal of Nutrition
Nutritional Epidemiology

See corresponding commentary on page 751.

Derivation and Validation of a Total Fruit and Vegetable Intake Prediction Model to Identify Targets for Biomarker Discovery Using the UK National Diet and Nutrition Survey

Elliot J Owen,[1,2] Sumaiya Patel,[1] Orla Flannery,[1] Tristan P Dew,[1,2,3] and Laura M O'Connor[1]

[1]Department of Health Professions, Faculty of Health, Psychology and Social Care, Manchester Metropolitan University, Manchester, United Kingdom; [2]Future Food Beacon of Excellence, University of Nottingham, Sutton Bonington, United Kingdom; and [3]School of Biosciences, University of Nottingham, Sutton Bonington, United Kingdom

ABSTRACT

Background: Dietary assessments in research and clinical settings are largely reliant on self-reported questionnaires. It is acknowledged that these are subject to measurement error and biases and that objective approaches would be beneficial. Dietary biomarkers have been purported as a complementary approach to improve the accuracy of dietary assessments. Tentative biomarkers have been identified for many individual fruits and vegetables (FVs), but an objective total FV intake assessment tool has not been established.

Objectives: To derive and validate a prediction model of total FV intake (TFVpred) to inform future biomarker studies.

Methods: Data from the National Diet and Nutrition Survey (NDNS) were used for this analysis. A modeling group (MG) consisting of participants aged >11 years from the NDNS years 5–6 was created (n = 1746). Intake data for 96 FVs were analyzed by stepwise regression to derive a model that satisfied 3 selection criteria: SEE ≤80, R^2 >0.7, and ≤10 predictors. The TFVpred model was validated using comparative data from a validation group (VG) created from the NDNS years 7–8 (n = 1865). Pearson's correlation coefficients were assessed between observed and predicted values in the MG and VG. Bland-Altman plots were used to assess agreement between TFVpred estimates and total FV intake.

Results: A TFVpred model, comprised of tomatoes, apples, carrots, bananas, pears, strawberries, and onions, satisfied the selection criteria (R^2 = 0.761; SEE = 78.81). Observed and predicted total FV intake values were positively correlated in the MG (r = 0.872; P < 0.001; R^2 = 0.761) and the VG (r = 0.838; P < 0.001; R^2 = 0.702). In the MG and VG, 95.0% and 94.9%, respectively, of TFVpred model residuals were within the limits of agreement.

Conclusions: Intakes of a concise FV list can be used to predict total FV intakes in a UK population. The individual FVs included in the TFVpred model present targets for biomarker discovery aimed at objectively assessing total FV intake.
J Nutr 2021;151:962–969.

Keywords: fruit and vegetables, prediction model, dietary assessment, biomarkers, dietary questionnaires

Introduction

Noncommunicable diseases (NCDs) accounted for 71.3% of worldwide mortality in 2016 (1). The objective measurement of modifiable risk factors is vital in informing strategies to reduce the public health burden incurred by NCDs. Fruit and vegetable (FV) intake has been associated with a lower risk of cardiovascular disease (2–5), type 2 diabetes (6, 7), and some forms of cancer (2, 8). These NCDs accounted for approximately 28.6 million deaths in 2016, equating to half of global mortality (1); thus, increasing FV consumption presents a potential opportunity to reduce the burden of disease.

Recent meta-analyses assessing the relationship between the quantity of FV intake and relative risk of all-cause mortality have produced equivocal results. Findings consistently indicate that the relative risk of all-cause mortality is proportionately lower with increased consumption of FVs, yet the reported plateau in risk reduction ranges from 5 servings (5) to 10 servings of FV per day (2). This 2-fold variation in the threshold of daily FV consumption at which there is the lowest relative risk of all-cause mortality is congruent with disparities in public health recommendations. The WHO and Public Health England currently recommend the consumption of at least 5 servings (400 g) of FV per day (9, 10), whereas the Danish Ministry of Food recommends the equivalent of 7.5 servings (600 g) per day (11). Findings from Aune et al. (2) infer that current recommendations, such as those of the United Kingdom presented in the Eatwell Guide (9), may not

© The Author(s) 2021. Published by Oxford University Press on behalf of the American Society for Nutrition. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

Manuscript received August 12, 2020. Initial review completed October 21, 2020. Revision accepted November 23, 2020.
First published online January 20, 2021; doi: https://doi.org/10.1093/jn/nxaa406.

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

sufficiently encourage the higher levels of FV consumption that pertain to a lower risk of all-cause mortality. The evidence regarding the optimal daily intake of FVs remains inconclusive, thus presenting a barrier toward informing public health recommendations, emphasizing the necessity for further elucidation of the relationship between FV intake and NCDs.

Epidemiological studies aiming to determine diet-disease relationships assess dietary intake using self-report methods, such as food diaries, 24-hour recalls, and FFQs (12–14). While necessary for obtaining data representative of habitual dietary intake, such methods are inherently subject to measurement error and biases and can be burdensome on participants (12, 15–17). A more succinct method of intake data collection—for example, reporting a single food group of interest—could alleviate the burden on participants, while conversely reducing the utility of the data when the exploration of whole diet–disease associations is required. Appropriate study designs and methodologies can mitigate the measurement error and biases inherent to self-report methods (18). A combined approach, comprised of the simultaneous measurement of dietary biomarkers and self-report methods, has been purported to improve the accuracy of dietary exposure measurements, thus facilitating the elucidation of diet-disease relations (18, 19).

Candidate dietary exposure biomarkers for the objective measurement of total FV intake, including carotenoids and polyphenols (20, 21), have been explored and were shown to have limited utility. The establishment of an objective tool to assess total FV intake, rather than individual FV intake, has not yet proved efficacious or been validated (22). Untargeted metabolomic techniques are increasingly prevalent within the literature, making significant progress in the identification and quantification of specific dietary exposure biomarkers (23, 24). The predominant focus of this research has been identifying single biomarkers for specific foods/food groups. Further to the identification of novel biomarkers, the use of a panel of biomarkers, by measuring a number of metabolites pertaining to a food/food group for a more accurate representation of dietary exposure, has been proposed (25). Multi-metabolite biomarker panels (MBPs) have been identified for the quantification of walnuts (26), bread (27), cocoa (28), orange juice (29), wine (30), and whole dietary patterns (31, 32); however, a panel for total FV intake has yet to be established.

The National Diet and Nutrition Survey (NDNS) is a continuous, cross-sectional survey designed to collect detailed quantitative information on the food consumption, nutrient intake, and nutritional status of the UK's general population (33). An analysis of these data can provide novel insight into total FV eating habits. The aim of this research was to identify a concise number of FVs that are predictive of total FV intake. Identifying such FVs stands to direct future metabolomic biomarker studies that pursue the objective measurement of FV intakes.

Funding for this project was provided by Manchester Metropolitan University.
Author disclosures: EJO, SP, OF, TPD, and LMO, no conflicts of interest.
Supplemental Table 1 and Supplemental Figure 1 are available from the "Supplementary data" link in the online posting of the article and from the same link in the table of contents at http://academic.oup.com/jn.
Address correspondence to EJO (e-mail: elliot.owen@mmu.ac.uk.).
Abbreviations used: AIC, Akaike information criterion; BIC, Bayesian information criterion; DINO, Diet In Nutrients Out; FFQ, food frequency questionnaire; FV, fruit and vegetable; LOA, limits of agreement; LR, likelihood ratio; MBP, multi-metabolite biomarker panel; MG, modeling group; NCD, noncommunicable disease; NDNS, National Diet and Nutrition Survey; TFVpred, total fruit and vegetable prediction; VG, validation group.

Methods

Study design

This study analyzed cross-sectional intake data of individuals from years 5–6 (2012/13 and 2013/14) and years 7–8 (2014/15 and 2015/16) of the NDNS rolling program (33, 34). The modeling data set (years 5–6) and validation data set (years 7–8) were retrieved from the UK data archive in September 2017 and January 2019, respectively.

Data source

Full methodological details of the NDNS have been described elsewhere (35). In short, the full NDNS data set from years 5–6 was comprised of 2546 participants (mean age, 30 ± 24 years) recruited from 323 postal-sector random-sampling units across the United Kingdom. Data were collected over 12 months to account for seasonal variation. Samples were stratified by country, ensuring proportional representation from England, Scotland, Wales, and Northern Ireland. Following initial interviews to obtain background information and familiarize participants with the intake data collection method, 4-day food diaries were completed and participants over the age of 4 years who consented to a nurse visit had anthropometric measurements (height, weight, waist and hip circumference, demi-span, blood pressure) and blood and urine samples taken. The modeling group (MG) data set was obtained from this sample and included all participants >11 years old ($n = 1746$).

Data processing

The faction of NDNS data used in the current analysis consisted of food and drink consumption data collected using 4-day unweighed food diaries (portions were quantified by household measures). Participants recorded the contents of all eating and drinking occasions over 4 consecutive days, including 1 weekend day. Food diaries were processed and coded using an adapted version of Health Nutrition Research's dietary assessment system, Diet In Nutrients Out (DINO) (36). DINO disaggregates composite items and items that differ by preparation into individual foods with a unique code. The current analysis aggregated data of the same fruit/vegetable with differing codes, to form daily intake values for individual FVs (g/day). Fruit juices, potatoes, and pulses (except for green beans, runner beans, and broad beans) were excluded from the analysis due to differences in nutrient composition from FV as included in the UK Eatwell Guide (9). We multiplied dried fruit intake by 3, based on the respective water and micronutrient content, to standardize dried and nondried FV intake (34). **Supplemental Table 1** outlines the details of individual FV intake data aggregation, FV consumption prevalences, and mean daily intakes in consumers only. Daily intakes of 96 FVs were calculated and used as potential predictor variables. Individual FV intakes were summed to calculate the total FV intake (g/day).

Statistical analysis

All data were obtained and processed using IBM SPSS Statistics 24 (SPSS, Inc.) and analyzed using Stata version 15 (StataCorp LLC). The assumptions of a multiple linear regression analysis were satisfied prior to analysis. Normality of residuals and homoscedasticity of the data were confirmed, and no transformations were applied to any variables. All potential predictors had a linear relationship with the total FV intake.

We conducted automated forward stepwise regression analyses. Models began with an intercept and were iteratively constructed by selecting the predictor variable (individual FV intake) that accounts for the most unique variance in the total FV intake. Subsequent models incorporated the individual fruit or vegetable that accounted for the most unique variance in total FV intake among the remaining predictor variables. Predictor variables were added with each model iteration

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

TABLE 1 Multiple linear regression models using individual fruit and vegetable intakes

Model	Predictor variables	Regression *P* value	Constant	Regression coefficient, β	R^2	Standard error of the estimate	Variance inflation factor
1	Tomatoes	<0.001	134.089	2.672	0.277	136.81	1.00
2	Tomatoes	<0.001	104.069	2.352	0.451	119.24	1.02
	Apples	<0.001	—	2.030	—	—	1.02
3	Tomatoes	<0.001	69.595	2.277	0.567	105.92	1.02
	Apples	<0.001	—	1.823	—	—	1.04
	Carrots	<0.001	—	2.982	—	—	1.02
4	Tomatoes	<0.001	46.973	2.091	0.664	93.26	1.04
	Apples	<0.001	—	1.546	—	—	1.07
	Carrots	<0.001	—	2.849	—	—	1.02
	Bananas	<0.001	—	1.406	—	—	1.06
5	Tomatoes	<0.001	45.125	2.060	0.702	87.91	1.04
	Apples	<0.001	—	1.452	—	—	1.08
	Carrots	<0.001	—	2.720	—	—	1.03
	Bananas	<0.001	—	1.292	—	—	1.08
	Pears	<0.001	—	1.362	—	—	1.05
6	Tomatoes	<0.001	39.892	1.995	0.732	83.33	1.04
	Apples	<0.001	—	1.453	—	—	1.08
	Carrots	<0.001	—	2.673	—	—	1.03
	Bananas	<0.001	—	1.250	—	—	1.08
	Pears	<0.001	—	1.391	—	—	1.05
	Strawberries	<0.001	—	1.762	—	—	1.01
7	Tomatoes	<0.001	29.877	1.773	0.761	78.81	1.11
	Apples	<0.001	—	1.428	—	—	1.08
	Carrots	<0.001	—	2.439	—	—	1.05
	Bananas	<0.001	—	1.211	—	—	1.08
	Pears	<0.001	—	1.422	—	—	1.05
	Strawberries	<0.001	—	1.714	—	—	1.01
	Onions	<0.001	—	1.519	—	—	1.11

Data from the National Diet and Nutrition Survey Rolling Program years 5–6 were used to predict total FV intake ($n = 1746$). Abbreviation: FV, fruit and vegetable.

until there was no longer an improvement in total FV intake variance accounted for by the model. Regression significance ($P < 0.05$) was taken to indicate that the independent variable predicts total FV intake. The variance inflation factor was used to quantify the correlation of predictors in a model, to detect any collinearity. Regression coefficients represent the mean change in an outcome for 1 unit of change in the predictor variable and were used to compile the regression equation. The SEE was calculated and R^2 was used to denote the proportion of variance in the total FV intake explained by each model.

Model selection criteria

The rationale underpinning model selection criteria was to produce a regression equation that could be used to facilitate the discovery of FV biomarkers. The future utility of the model is dependent upon having few predictors to moderate the extent of biomarker measurement required, while explaining a large proportion of the variance in predicted total FV intake. We established iterative models that satisfied 3 pragmatically determined selection criteria: having an SEE $\leq$ an 80 g FV serving; having variance in total FV intake (R^2) >0.7; and capping the number of predictors in the model at 10 to produce a concise assessment tool. A comparative assessment of regression models was facilitated by calculating the adjusted R^2, Akaike information criterion (AIC), Bayesian information criterion (BIC), and penalized likelihood ratio (LR). The aim of all comparative assessments was to ensure that all subsequent models were an improvement on the previous model.

Model validation

Validation of the final total FV prediction model iteration (TFVpred) was conducted using a novel data set from the NDNS years 7–8, with participants aged >11 years. NDNS data collection methodologies were consistent with the years 5–6 used as the MG. The current analysis applied the same data processing procedure described above to the validation group (VG) data set to obtain comparable FV intake data. The TFVpred equation was applied to the VG data set to predict total FV intake (g/day). Pearson's r correlation coefficient was measured to determine linearity between observed and predicted total FV values. R^2 was calculated to measure the amount of variance in the TFVpred estimated total FV intake explained by the observed total FV intake. A correlational analysis was conducted with observed and predicted FV intakes in vegetarian and vegan subsets of the MG and VG to assess the validity of the prediction model in a subset of the population with known differences in FV consumption patterns. Bland-Altman plots were generated to assess the agreement between TFVpred estimates and observed total FV intakes in modeling and validation groups. Limits of agreement were plotted at $\pm$1.96 SDs of the mean difference between the observed and predicted values of total FV intake.

Results

Multiple linear regression models for prediction of total FV intake

In total, 4-day food diaries were analyzed from 1746 participants in the MG and 1865 participants in the VG. Forward stepwise regression model summaries are displayed in Table 1. Total FV prediction Model 7 (TFVpred) was the first model iterated that met all model selection criteria, with an R^2 >0.7, an SEE <80, and $\leq$10 predictor variables. All 7 models predicted total FV intake ($P < 0.05$). The proportion of variance explained by regression models (R^2) increased from 0.277 to 0.761 between Models 1 and 7. Incremental reductions in SEE were observed with each regression model including a novel

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

TABLE 2 Comparison of multiple linear regression models using individual fruit and vegetable intakes

Model	Cumulative predictor variables	Adjusted R^2	Change in adjusted R^2	Akaike information criterion	Bayesian information criterion	LR models tested	LR test statistic	LR test *P*
1	Tomatoes	0.276	—	22133	22144	—	—	
2	Apples	0.450	0.174	21654	21670	1 and 2	481.13	<0.001
3	Carrots	0.566	0.116	21241	21263	2 and 3	414.65	<0.001
4	Bananas	0.664	0.098	20798	20825	3 and 4	445.38	<0.001
5	Pears	0.701	0.037	20592	20625	4 and 5	207.35	<0.001
6	Strawberries	0.732	0.031	20406	20445	5 and 6	187.86	<0.001
7	Onions	0.760	0.028	20213	20256	6 and 7	195.84	<0.001

Data from the National Diet and Nutrition Survey Rolling Program years 5–6 were used to predict total FV intake ($n = 1746$). Abbreviations: FV, fruit and vegetable; LR, likelihood ratio.

predictor. TFVpred, comprised of 7 predictor FV coefficients and a constant, is displayed in **Equation 1**:

$$\begin{aligned}\textbf{TFVpred} = {} & 1.773\ (\textbf{tomatoes}) + 1.428\ (\textbf{apples}) \\ & + 2.439\ (\textbf{carrots}) + 1.211\ (\textbf{bananas}) \\ & + 1.422\ (\textbf{pears}) + 1.714\ (\textbf{strawberries}) \\ & + 1.519\ (\textbf{onions}) + 29.88\ (\textbf{constant}) \qquad (1)\end{aligned}$$

The TFVpred equation highlights the 7 predictor FVs accounting for the most variance in total FV intake—namely, tomatoes, apples, carrots, bananas, pears, strawberries, and onions—thus presenting targets for intake biomarker discovery. There were 5 FVs included in the TFVpred model (tomatoes, onions, carrots, bananas, and apples) that were within the top 6 most commonly consumed FVs (as per number of consumers) in the MG, while strawberries and pears were numbers 15 and 20, respectively (Supplemental Table 1). All predictor variable FVs were within the top 40 FVs for mean daily intakes in consumers only.

Model comparison

A comparison of regression models is shown in **Table 2**. The variance in total FV intake explained by the models, when corrected for the number of predictors, incrementally increased with an additional model iteration. The size of incremental augmentation in adjusted R^2 diminished as the regression models progressed, with the maximum change being an increase of 0.174 from Model 1 to Model 2 and the smallest change being 0.028 from Model 6 to Model 7. The penalized-LR criteria, AIC, and BIC are presented for each model in Table 2. AIC and BIC values were incrementally smaller as more predictors were added to the regression models. LR tests for nested models were significant with all subsequent iterations, indicating successive improvements in goodness of fit.

Model validation

In the MG, the observed and predicted values of total FV intake were positively correlated (r = 0.872; $P < 0.001$), with an R^2 of 0.761 (**Figure 1**A). Observed and predicted total FV intake values in the VG were also positively correlated (r = 0.838;

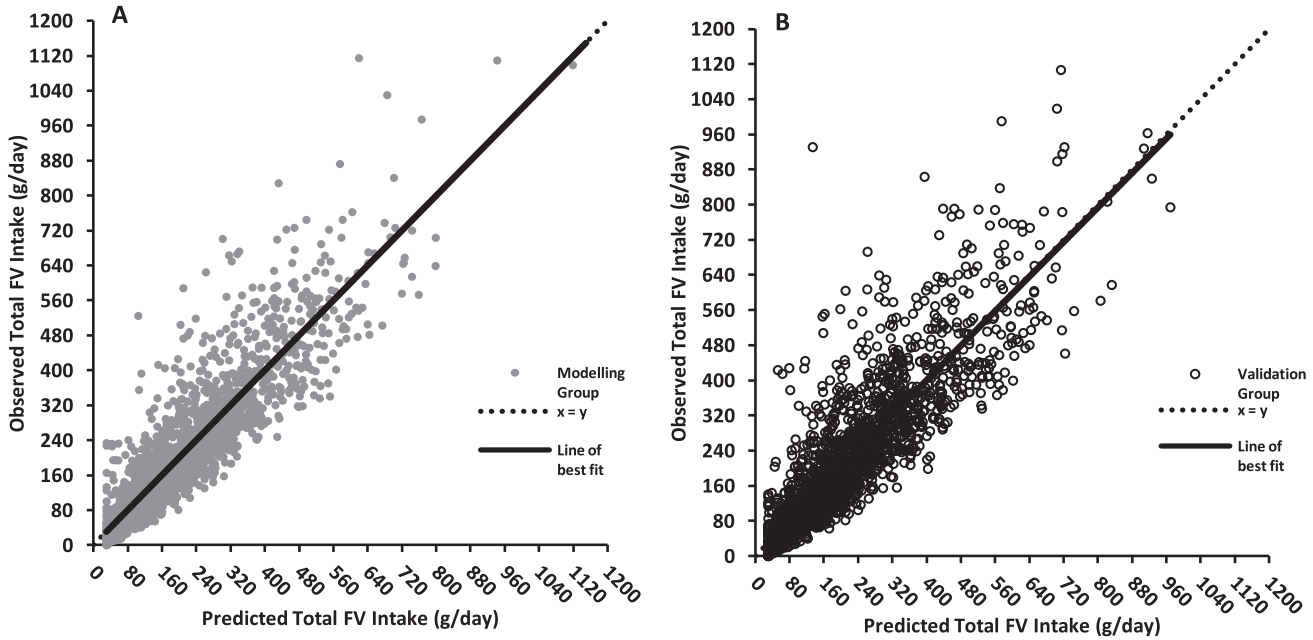

FIGURE 1 Correlation between observed and predicted total FV intake using the TFVpred equation for the (A) modeling group (NDNS years 5–6; $n = 1746$) and (B) validation group (NDNS years 7–8; $n = 1865$). Abbreviations: FV, fruit and vegetable; NDNS, National Diet and Nutrition Survey; TFVpred, total fruit and vegetable prediction.

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

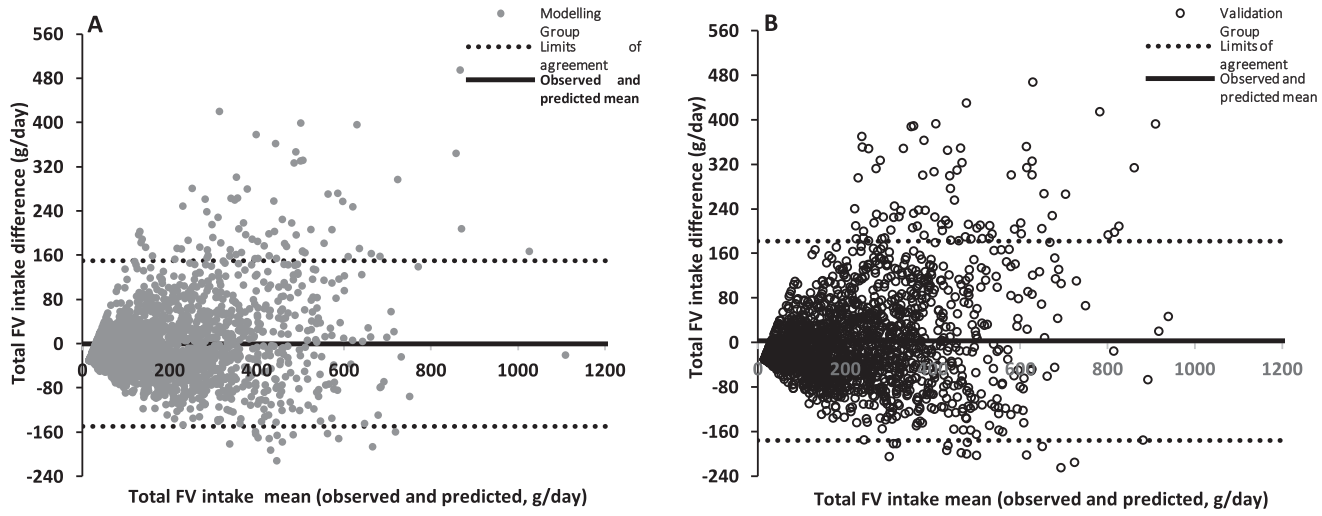

FIGURE 2 Bland-Altman plots of total FV intake predictions in the (A) modeling group ($n = 1746$) and (B) validation group ($n = 1865$). Plots display the difference between total FV intake measured by the NDNS and total FV intake predicted by the TFVpred model versus the observed and predicted mean. Limits of agreement (dotted lines) are displayed at ±1.96 SDs of the mean difference between the observed and predicted values of total FV intake. Abbreviations: FV, fruit and vegetable; NDNS, National Diet and Nutrition Survey; TFVpred, total fruit and vegetable prediction.

$P < 0.001$), with an R^2 of 0.702 (Figure 1B). Bland-Altman plots determined there was good agreement between observed and predicted total FV intake values, with the MG (**Figure 2**A) and VG (Figure 2B) demonstrating 95.0% and 94.9%, respectively, of residuals were within the limits of agreement. Observed and predicted total FV intake values within vegetarian and vegan subsets were positively correlated in both the MG (r = 0.882; $P < 0.001$; $R^2 = 0.777$; **Supplemental Figure 1A**) and VG (r = 0.839; $P < 0.001$; $R^2 = 0.704$; Supplemental Figure 1B).

Discussion

To our knowledge, this is the first study to elucidate a concise group of individual FVs that are predictive of total FV intake, accounting for 76.1% of total variance. The seventh model iteration, TFVpred, was the first to satisfy the predetermined selection criteria and was subsequently used to predict the total FV intake in the VG, using individual intake values of tomatoes, apples, carrots, bananas, pears, strawberries, and onions. A correlational analysis and Bland-Altman plots were used to assess the efficacy of the TFVpred model when applied to the VG, and demonstrated strong agreement between observed and predicted values. TFVpred thus provides a potential assessment tool in estimating the total FV intake where valid measurements of 7 individual FV intakes (tomatoes, apples, carrots, bananas, pears, strawberries, and onions) are available. A multitude of comparisons between models were conducted to determine that TFVpred outperforms other models by AIC, BIC, and LR test statistics, and is therefore the most appropriate model for estimating total FV intake (37). This research has the potential to consolidate the applicability of existing individual FV measurements obtained using dietary questionnaires. Furthermore, the identified FVs signify clear targets for novel biomarker discovery. The subsequent integration of validated biomarkers within the TFVpred equation can provide additional utility as a potential tool for total FV intake estimation.

Dietary questionnaires

Self-report methods of dietary intake assessment, such as food diaries, 24-hour recalls, and FFQs, have been a longstanding topic of debate in nutritional research (17, 38), while remaining the most prevalent techniques to assess diet-disease relationships (4, 39). Critics state that the reliance on memory and the influence of researcher/social-approval biases can incur random and systematic measurement errors, such as the over-reporting of FV intake (12–14, 17). Furthermore, the accuracy of self-reported data may be influenced by the ability of individuals to quantify the size and contents of a FV serving, or by the sensitivity of the assessment method (40, 41). Proponents of self-report methods acknowledge that while limitations exist, study design considerations and corrections for measurement error can be applied to gather insightful intake data that are currently unobtainable using other means (42, 43). The NDNS data set used in the current study aimed to collect data accurately pertaining to the UK population by mitigating the effect of some of these limitations through an appropriate study design. Daily food diaries were completed over 4 consecutive days to minimize reliance on memory (42). Upon completion of food diaries, trained interviewers met with participants to aid the quantification of the food diary constituents, where original visual aids were insufficient (35). The NDNS data set presents a useful source when compiling inferential statistical models, as in the present analysis. Given the robustness of the NDNS methodology, validation with an updated NDNS data set was necessary and demonstrated the efficacy of the TFVpred model as a practical tool for total FV intake estimation.

The novel assessment of total FV intake using the TFVpred model could utilize existing methods of measuring individual FV intakes from dietary questionnaires. Measurements could be obtained via amended FFQs—for instance, FFQs condensed to include only an FV assessment—providing sufficient validation is conducted (39, 44, 45). Kristjansdottir et al. (44) reported that FV intake estimated using a combined 24-hour recall and an FFQ was associated with 7-day food diary reported intake, with a Spearman's coefficient of 0.73 ($P < 0.001$). Furthermore, Block et al. (46) correlated FV intakes obtained using 100-item

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

FFQs (47) and a single page screener questionnaire, reporting a Spearman's coefficient of 0.71 ($P < 0.001$). Using a screener to assess FV intake could provide a time-effective alternative to a lengthy questionnaire and provide specific FV intake data. A practical application of the predictive FVs identified in the present analysis would be to incorporate these FVs in screener questionnaires or as prompts in multiple-pass dietary assessment methods. Adopting such changes may increase the accuracy of dietary intake data, though amendments to validated dietary assessment tools would require subsequent validation. Incorporating measurements of the FVs identified in the TFVpred model within existing dietary questionnaires presents an inexpensive tool for internal validation to improve the precision of dietary intake assessments.

Combining dietary questionnaires and biomarkers

The prevailing recommendations from prominent research groups within the field of nutrition and dietary assessment include the combined assessment of diet using dietary questionnaires and biomarker quantification (18, 19, 25). A prospective application of the TFVpred model validated in the present analysis would be to integrate biomarker assessments for the 7 FVs, providing an objective assessment tool that can be obtained from biological samples and be used to assess FV exposure alongside appropriately conducted questionnaires. The NDNS represents an example of how this may be achieved, due to the concurrent collection of self-report data and urine samples; however, a FV biomarker assessment panel has yet to be established and validated (35). Systematic reviews exploring the efficacy of objective assessments of FV intake by dose-dependent concentration biomarkers have ascertained that no single candidate biomarker can accurately measure total FV intake (20, 48). However, putative dose-dependent urinary biomarkers have been identified for some FVs, including grapes (49), peas, apples, onions (50), red cabbage, strawberries, and beetroot (31). Prevalent techniques aiming to identify a panel of biomarkers pertaining to individual foods/food groups include targeted and untargeted tandem high-performance LC-MS, as well as proton nuclear magnetic resonance spectroscopy, with subsequent multivariate modeling (Principal Component-Discriminant Analysis, Partial Least Squares, and Random Forest Classification) (27, 32, 51). This has led to the identification of numerous metabolites purported as biomarkers of dietary exposure, although validation as dose-dependent biomarkers of intake, which would be necessary prior to TFVpred model integration, is less pervasive (49, 52, 53). The specificity of putative biomarkers ranges from identifying individual foods (including FVs) to broad dietary patterns (32, 54, 55).

Potentially confounding factors for biomarker identification include inherent genetic variance between individuals, physiological and lifestyle factors that may influence metabolism, biological sample handling, and the analytical methodology (22). Future research should aim to negate some of these factors. For example, Garcia-Aloy et al. (25) propose the use of MBPs to provide insight into dietary exposure. MBPs enable the simultaneous measurement of numerous metabolites that pertain to a specific food/food group, capturing a broader faction of dietary exposure. Once validated, prospective MBPs of individual FV intakes could be integrated with the regression equation modeled in the present study as a method of estimating the total FV intake. Dragsted et al. (56) identified a stringent set of post-discovery validity criteria for biomarkers, including assessments of: *1*) biochemical plausibility and stability; *2*) saturation kinetics and dose-dependency with low abundancy when intake is 0; *3*) time-responsiveness to inform when biological samples can be collected; *4*) robustness after co-ingestion with other foods; *5*) reliability to ensure biomarkers are comparable to assessments from other questionnaire or biomarker measurements; and *6*) a reproducible analytical methodology. Meeting these standards is imperative if biomarkers are to improve the precision and accuracy of dietary assessments. Considerable work is necessary to elucidate, in particular, time-responsiveness and dose-dependency of putative FV biomarkers (25). At present, the limitations associated with both facets of dietary assessment cannot be fully alleviated by adopting sole usage of the alternate technique; thus, combinations of dietary questionnaires and biomarker assessments should be explored (16, 25).

Strengths and limitations

FV servings of 80 g were used in the present analysis to compute regression models; thus, FVs that deviated from the standard 80-g serving sizes, such as dried fruits, required numerical transformation prior to being considered a FV portion. This was conducted to prevent the potential exclusion of a subset of FVs that contribute to total FV intake, but do not constitute a regular FV serving. Some semi-dried fruits were not included in the current analysis due to the unknown composition of portion sizes. Consistent with other nutritional epidemiology research (57, 58), children aged <12 years (MG, $n = 763$; VG, $n = 822$) were excluded from the current analysis to mitigate the systematic error incurred by having dissimilar eating trends and serving sizes for adolescents and adults. As the current analysis was conducted using intake data from UK-based participants ≥12 years old, the TFVpred model should not be prospectively used to estimate total FV intakes in children <12 years old. Deriving the TFVpred model using stepwise linear regression modeling and pragmatic predetermined selection criteria facilitated the formation of a model that included a combination of influential FVs that were predictive of total FV intake and were frequently consumed in the population. TFVpred predictor FVs were among the most pervasively consumed in the MG and VG, indicating good suitability within a UK population. Future research should investigate the efficacy of the TFVpred model in other developed countries, and further validation is required prior to use in populations outside of the United Kingdom, as FV intakes are variable between countries (59, 60). A prominent challenge within the present study was producing a model with a small number of predictors that captured a substantial proportion of the variance in total FV intake without including relevant cofactors, such as socioeconomic status (61, 62), food availability (63), and vegetarianism (64). The TFVpred model predictions were accurate for subsets of the population known to have different FV consumption patterns, as demonstrated by the correlation between observed and predicted total FV intakes in vegetarians and vegans. The TFVpred model also performed well across a broad variety of FV intakes from the small proportion of individuals that fall outside the upper limits of agreement (LOA). Bland-Altman plots (Figure 2) indicate that 4.70% and 4.86% of individuals in the MG and VG, respectively, fall outside the upper LOA, thus consuming a variety of FVs that are not accounted for by the model. The simultaneous assessment of cofactors of total FV intake and additional FVs would increase the accuracy of prediction models; however, the aim of the present study was to identify a concise number of predictor FVs that can be integrated into dietary questionnaires to reliably

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

estimate total FV intakes in a UK population and identify targets for biomarker discovery, rather than to establish a multifaceted prediction model of total FV intake.

Conclusions

The TFVpred model (Equation 1) established in the current study provides a valuable tool for estimating total FV intake. Future utility of the TFVpred model would be improved with the integration of dose-dependent biomarkers/MBPs for those FVs that predict total FV intake (tomatoes, apples, carrots, bananas, pears, strawberries, and onions). The identification of these FVs through the establishment and validation of the TFVpred model provides a clear pathway for future research by identifying dose-dependent biomarker targets. Advances in biomarker identification and validation provide a valuable opportunity to obtain objective assessments of total FV intake that, in parallel with appropriate self-report techniques, could denote notable improvements in the accuracy of dietary assessments.

Acknowledgments

The authors' responsibilities were as follows—EJO, TPD, and LMO'C: developed the research question and planned the analysis; EJO: performed the analysis and drafted the manuscript; SP and LMO'C: provided analysis support; OF, TPD, and LMO'C: provided editorial support for the manuscript; and all authors: read and approved the final manuscript.

References

1. World Health Organization. Global health estimates 2016: deaths by cause, age, sex, by country and by region, 2000–2016 [Internet]. Geneva (Switzerland): WHO; 2018. [Last accessed 2020 Aug 11]. Available from: https://www.who.int/healthinfo/global_burden_disease/estimates/en/
2. Aune D, Giovannucci E, Boffetta P, Fadnes LT, Keum N, Norat T, Greenwood DC, Riboli E, Vatten LJ, Tonstad S. Fruit and vegetable intake and the risk of cardiovascular disease, total cancer and all-cause mortality–a systematic review and dose-response meta-analysis of prospective studies. Int J Epidemiol 2017;46:1029–56.
3. Boeing H, Bechthold A, Bub A, Ellinger S, Haller D, Kroke A, Leschik-Bonnet E, Müller MJ, Oberritter H, Schulze M, et al. Critical review: vegetables and fruit in the prevention of chronic diseases. Eur J Nutr 2012;51:637–63.
4. Crowe FL, Roddam AW, Key TJ, Appleby PN, Overvad K, Jakobsen MU, Tjønneland A, Hansen L, Boeing H, Weikert C, et al. Fruit and vegetable intake and mortality from ischaemic heart disease: results from the European Prospective Investigation into Cancer and Nutrition (EPIC)-Heart study. Eur Heart J 2011;32:1235–43.
5. Wang X, Ouyang Y, Liu J, Zhu M, Zhao G, Bao W, Hu FB. Fruit and vegetable consumption and mortality from all causes, cardiovascular disease, and cancer: systematic review and dose-response meta-analysis of prospective cohort studies. BMJ 2014;349:g4490. doi: 10.1136/bmj.g4490.
6. Cooper AJ, Forouhi NG, Ye Z, Buijsse B, Arriola L, Balkau B, Barricarte A, Beulens JW, Boeing H, Büchner FL, et al. Fruit and vegetable intake and type 2 diabetes: EPIC-InterAct prospective study and meta-analysis. Eur J Clin Nutr 2012;66:1082–92.
7. Cooper AJ, Sharp SJ, Lentjes MA, Luben RN, Khaw K-T, Wareham NJ, Forouhi NG. A prospective study of the association between quantity and variety of fruit and vegetable intake and incident type 2 diabetes. Diabetes Care 2012;35:1293–300.
8. Boffetta P, Couto E, Wichmann J, Ferrari P, Trichopoulos D, Bueno-de-Mesquita HB, Van Duijnhoven FJ, Büchner FL, Key T, Boeing H, et al. Fruit and vegetable intake and overall cancer risk in the European Prospective Investigation into Cancer and Nutrition (EPIC). J Natl Cancer Inst 2010;102:529–37.
9. The Eatwell Guide [Internet]. 2018. [Last accessed 2020 Aug 11]. Available from: https://www.nhs.uk/live-well/eat-well/the-eatwell-guide/
10. WHO. Increasing fruit and vegetable consumption to reduce the risk of noncommunicable diseases [Internet]. Geneva (Switzerland): WHO. [Last accessed 2020 Aug 11]. Available from: http://www.who.int/elena/titles/fruit_vegetables_ncds/en/.
11. Tetens I, Andersen LB, Astrup A, Gondolf UH, Hermansen K, Jakobsen MU, Knudsen VK, Mother H, Schwartz P, Tjønneland A, et al. Evidence basis for Danish advice on diet and physical activity. (Danish title: Evidensgrundlaget for danske råd om kost og fysisk aktivitet. DTU Fødevareinstituttet.). DTU Food Inst Søborg (Denmark): 2013. p. 164.
12. Thompson FE, Subar AF. Chapter 1–Dietary assessment methodology. In: Coulston AM, Boushey CJ, Ferruzzi MG, Delahanty LM, editors. Nutrition in the prevention and treatment of disease. (Fourth Edition). Cambridge (MA): Academic Press; 2017. pp. 5–48.
13. Hebert JR, Hurley TG, Peterson KE, Resnicow K, Thompson FE, Yaroch AL, Ehlers M, Midthune D, Williams GC, Greene GW, et al. Social desirability trait influences on self-reported dietary measures among diverse participants in a multicenter multiple risk factor trial. J Nutr 2008;138:226S–34S.
14. Miller TM, Abdel-Maksoud MF, Crane LA, Marcus AC, Byers TE. Effects of social approval bias on self-reported fruit and vegetable consumption: a randomized controlled trial. Nutr J 2008;7:18. https://doi.org/10.1186/1475-2891-7-18.
15. Freedman LS, Schatzkin A, Wax Y. The impact of dietary measurement error on planning sample size required in a cohort study. Am J Epidemiol 1990;132:1185–95.
16. Freedman LS, Schatzkin A, Midthune D, Kipnis V. Dealing with dietary measurement error in nutritional cohort studies. J Natl Cancer Inst 2011;103:1086–92.
17. Archer E, Marlow ML, Lavie CJ. Controversy and debate: memory-based methods paper 1: the fatal flaws of food frequency questionnaires and other memory-based dietary assessment methods. J Clin Epidemiol 2018;104:113–24.
18. 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;145:2639–45.
19. Brouwer-Brolsma EM, Brennan L, Drevon CA, van Kranen H, Manach C, Dragsted LO, Roche HM, Andres-Lacueva C, Bakker SJL, Bouwman J, et al. Combining traditional dietary assessment methods with novel metabolomics techniques: present efforts by the Food Biomarker Alliance. Proc Nutr Soc 2017;76:619–27.
20. Baldrick FR, Woodside JV, Elborn JS, Young IS, McKinley MC. Biomarkers of fruit and vegetable intake in human intervention studies: a systematic review. Crit Rev Food Sci Nutr 2011;51:795–815.
21. Woodside JV, Draper J, Lloyd A, McKinley MC. Use of biomarkers to assess fruit and vegetable intake. Proc Nutr Soc 2017;76:308–15.
22. Scalbert A, Brennan L, Manach C, Andres-Lacueva C, Dragsted LO, Draper J, Rappaport SM, van der Hooft JJ, Wishart DS. The food metabolome: a window over dietary exposure. Am J Clin Nutr 2014;99:1286–308.
23. Guasch-Ferré M, Bhupathiraju SN, Hu FB. Use of metabolomics in improving assessment of dietary intake. Clin Chem 2018;64:82–98.
24. Collins C, McNamara AE, Brennan L. Role of metabolomics in identification of biomarkers related to food intake. Proc Nutr Soc 2019;78:189–96.
25. Garcia-Aloy M, Rabassa M, Casas-Agustench P, Hidalgo-Liberona N, Llorach R, Andres-Lacueva C. Novel strategies for improving dietary exposure assessment: multiple-data fusion is a more accurate measure than the traditional single-biomarker approach. Trends Food Sci Technol 2017;69:220–9.
26. Garcia-Aloy M, Llorach R, Urpi-Sarda M, Tulipani S, Estruch R, Martínez-González MA, Corella D, Fitó M, Ros E, Salas-Salvadó J, et al. Novel multimetabolite prediction of walnut consumption by a urinary biomarker model in a free-living population: the PREDIMED Study. J Proteome Res 2014;13:3476–83.
27. Garcia-Aloy M, Llorach R, Urpi-Sarda M, Tulipani S, Salas-Salvadó J, Martínez-González MA, Corella D, Fitó M, Estruch R, Serra-Majem L, et al. Nutrimetabolomics fingerprinting to identify biomarkers of bread exposure in a free-living population from the PREDIMED study cohort. Metabolomics 2015;11:155–65.
28. Garcia-Aloy M, Llorach R, Urpi-Sarda M, Jáuregui O, Corella D, Ruiz-Canela M, Salas-Salvadó J, Fitó M, Ros E, Estruch R, et al. A metabolomics-driven approach to predict cocoa product consumption

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022

by designing a multimetabolite biomarker model in free-living subjects from the PREDIMED study. Mol Nutr Food Res 2015;59:212–20.
29. Rangel-Huerta OD, Aguilera CM, Perez-de-la-Cruz A, Vallejo F, Tomas-Barberan F, Gil A, Mesa MD. A serum metabolomics-driven approach predicts orange juice consumption and its impact on oxidative stress and inflammation in subjects from the BIONAOS study. Mol Nutr Food Res 2017;61:1600120. doi: 10.1002/mnfr.201600120.
30. Vázquez-Fresno R, Llorach R, Urpi-Sarda M, Khymenets O, Bulló M, Corella D, Fitó M, Martínez-González MA, Estruch R, Andres-Lacueva C. An NMR metabolomics approach reveals a combined-biomarkers model in a wine interventional trial with validation in free-living individuals of the PREDIMED study. Metabolomics 2015;11:797–806.
31. Andersen M-BS, Kristensen M, Manach C, Pujos-Guillot E, Poulsen SK, Larsen TM, Astrup A, Dragsted L. Discovery and validation of urinary exposure markers for different plant foods by untargeted metabolomics. Anal Bioanal Chem 2014;406:1829–44.
32. Lloyd AJ, Willis ND, Wilson T, Zubair H, Xie L, Chambers E, Garcia-Perez I, Tailliart K, Beckmann M, Mathers JC, et al. Developing a food exposure and urine sampling strategy for dietary exposure biomarker validation in free-living individuals. Mol Nutr Food Res 2019;63:1900062. https://doi.org/10.1002/mnfr.201900062.
33. NatCen Social Research. National diet and nutrition survey years 1–6, 2008/09–2013/14 [Internet]. 2017. [Last accessed 2020 Aug 11]. Available from: 10.5255/UKDA-SN-6533-7.
34. MRC Elsie Widdowson Laboratory. National Diet and Nutrition Survey years 1–8, 2008/09–2015/16 [Internet]. 2018. [Last accessed 2020 Aug 11]. Available from: 10.5255/UKDA-SN-6533-11.
35. Bates B, Cox L, Nicholson S, Page P, Prentice A, Steer T, Swan G, editors. National Diet and Nutrition Survey. Results from years 5 and 6 (combined) of the rolling programme (2012/2013–2013/2014). [Internet], London (UK): Public Health England; 2016. [Last accessed 2020 Aug 11]. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/551352/NDNS_Y5_6_UK_Main_Text.pdf.
36. Fitt E, Mak TN, Stephen AM, Prynne C, Roberts C, Swan G, Farron-Wilson M. Disaggregating composite food codes in the UK National Diet and Nutrition Survey food composition databank. Eur J Clin Nutr 2010;64:S32–6.
37. Dziak JJ, Coffman DL, Lanza ST, Li R, Jermiin LS. Sensitivity and specificity of information criteria. Brief Bioinform 2019;21:553–65.
38. Martín-Calvo N, Martínez-González MÁ. Controversy and debate: memory-based dietary assessment methods paper 2. J Clin Epidemiol Elmsford 2018;104:125–9.
39. Miller V, Mente A, Dehghan M, Rangarajan S, Zhang X, Swaminathan S, Dagenais G, Gupta R, Mohan V, Lear S, et al. Fruit, vegetable, and legume intake, and cardiovascular disease and deaths in 18 countries (PURE): a prospective cohort study. Lancet North Am Ed 2017;390:2037–49.
40. Rooney C, McKinley MC, Appleton KM, Young IS, McGrath AJ, Draffin CR, Hamill LL, Woodside JV. How much is "5-a-day"? A qualitative investigation into consumer understanding of fruit and vegetable intake guidelines. J Hum Nutr Diet 2017;30:105–13.
41. Roark RA, Niederhauser VP. Fruit and vegetable intake: issues with definition and measurement. Public Health Nutr 2013;16:2–7.
42. Shim J-S, Oh K, Kim HC. Dietary assessment methods in epidemiologic studies. Epidemiol Health 2014;36:e2014009.
43. Bennett DA, Landry D, Little J, Minelli C. Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. BMC Med Res Methodol 2017;17:146. https://doi.org/10.1186/s12874-017-0421-6.
44. Kristjansdottir AG, Andersen LF, Haraldsdottir J, Almeida MDV de, Thorsdottir I. Validity of a questionnaire to assess fruit and vegetable intake in adults. Eur J Clin Nutr 2006;60:408–15.
45. Penkilo M, George GC, Hoelscher DM. Reproducibility of the School-Based Nutrition Monitoring Questionnaire among fourth-grade students in Texas. J Nutr Educ Behav 2008;40:20–7.
46. Block G, Gillespie C, Rosenbaum EH, Jenson C. A rapid food screener to assess fat and fruit and vegetable intake. Am J Prev Med 2000;18:284–8.
47. 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;124:453–69.
48. Pennant M, Steur M, Moore C, Butterworth A, Johnson L. Comparative validity of vitamin C and carotenoids as indicators of fruit and vegetable intake: a systematic review and meta-analysis of randomised controlled trials. Br J Nutr 2015;114:1331–40.
49. Garcia-Perez I, Posma JM, Chambers ES, Nicholson JK, Mathers JC, Beckmann M, Draper J, Holmes E, Frost G. An analytical pipeline for quantitative characterization of dietary intake: application to assess grape intake. J Agric Food Chem 2016;64:2423–31.
50. Posma JM, Garcia-Perez I, Heaton JC, Burdisso P, Mathers JC, Draper J, Lewis M, Lindon JC, Frost G, Holmes E, et al. Integrated analytical and statistical two-dimensional spectroscopy strategy for metabolite identification: application to dietary biomarkers. Anal Chem 2017;89:3300–9.
51. Garcia-Perez I, Posma JM, Gibson R, Chambers ES, Hansen TH, Vestergaard H, Hansen T, Beckmann M, Pedersen O, Elliott P, et al. Objective assessment of dietary patterns by use of metabolic phenotyping: a randomised, controlled, crossover trial. Lancet Diabetes Endocrinol 2017;5:184–95.
52. Ulaszewska M, Vázquez-Manjarrez N, Garcia-Aloy M, Llorach R, Mattivi F, Dragsted LO, Praticò G, Manach C. Food intake biomarkers for apple, pear, and stone fruit. Genes Nutr 2018;13:29. doi: 10.1186/s12263-018-0620-8.
53. Gibbons H, Michielsen CJR, Rundle M, Frost G, McNulty BA, Nugent AP, Walton J, Flynn A, Gibney MJ, Brennan L. Demonstration of the utility of biomarkers for dietary intake assessment; proline betaine as an example. Mol Nutr Food Res 2017;61:10. doi: 10.1002/mnfr.201700037.
54. Edmands WM, Beckonert OP, Stella C, Campbell A, Lake BG, Lindon JC, Holmes E, Gooderham NJ. Identification of human urinary biomarkers of cruciferous vegetable consumption by metabonomic profiling. J Proteome Res 2011;10:4513–21.
55. Landberg R, Hanhineva K. Biomarkers of a healthy Nordic diet–from dietary exposure biomarkers to microbiota signatures in the metabolome. Nutrients 2020;12:27. https://doi.org/10.3390/nu12010027.
56. Dragsted LO, Gao Q, Scalbert A, Vergères G, Kolehmainen M, Manach C, Brennan L, Afman LA, Wishart DS, Andres Lacueva C, et al. Validation of biomarkers of food intake–critical assessment of candidate biomarkers. Genes Nutr 2018;13:14. https://doi.org/10.1186/s12263-018-0603-9.
57. Jones NR, Tong TY, Monsivais P. Meeting UK dietary recommendations is associated with higher estimated consumer food costs: an analysis using the National Diet and Nutrition Survey and consumer expenditure data, 2008–2012. Public Health Nutr 2018;21:948–56.
58. Dikariyanto V, Berry SE, Pot GK, Francis L, Smith L, Hall WL. Tree nut snack consumption is associated with better diet quality and CVD risk in the UK adult population: National Diet and Nutrition Survey (NDNS) 2008–2014. Public Health Nutr 2020;23:3160–69.
59. Stea TH, Nordheim O, Bere E, Stornes P, Eikemo TA. Fruit and vegetable consumption in Europe according to gender, educational attainment and regional affiliation–a cross-sectional study in 21 European countries. PLoS One 2020;15:e0232521.
60. Ray S, Nicholson S, Ziauddeen N, Steer T, Cole D, Solis-Trapala I, Amoutzopoulos B, Page P. What do we know about fruit and vegetable consumption in the UK? Trends from the National Diet and Nutrition Survey Rolling Programme (NDNS RP). FASEB J 2015;29:LB407.
61. Mak TN, Prynne CJ, Cole D, Fitt E, Bates B, Stephen AM. Patterns of sociodemographic and food practice characteristics in relation to fruit and vegetable consumption in children: results from the UK National Diet and Nutrition Survey Rolling Programme (2008–2010). Public Health Nutr 2013;16:1912–23.
62. Yau A, Adams J, Monsivais P. OP48 age, sex and socioeconomic inequalities in fruit and vegetable intake in UK adults, 1986–2012. J Epidemiol Community Health 2017;71:A24–5.
63. Kamphuis CBM, Giskes K, de Bruijn G-J, Wendel-Vos W, Brug J, van Lenthe FJ. Environmental determinants of fruit and vegetable consumption among adults: a systematic review. Br J Nutr 2006;96:620–35.
64. Pollard J, Greenwood D, Kirk S, Cade J. Lifestyle factors affecting fruit and vegetable consumption in the UK Women's Cohort Study. Appetite 2001;37:71–9.

Downloaded from https://academic.oup.com/jn/article/151/4/962/6109677 by University of Nottingham user on 28 June 2022