HHS Public Access

Author manuscript

Eur J Nutr. Author manuscript; available in PMC 2022 December 01.

Published in final edited form as:

Eur J Nutr. 2021 December ; 60(8): 4207–4218. doi:10.1007/s00394-021-02577-1.

Evaluation of potential metabolomics-based biomarkers of protein, carbohydrate and fat intakes using a controlled feeding study

Cheng Zheng[1], **G.A. Nagana Gowda**[2], **Daniel Raftery**[2,3], **Marian L. Neuhouser**[3], **Lesley F. Tinker**[3], **Ross L. Prentice**[3], **Shirley A.A. Beresford**[3,4], **Yiwen Zhang**[5], **Lisa Bettcher**[2], **Robert Pepin**[2], **Danijel Djukovic**[2], **Haiwei Gu**[2], **Gregory A. Barding Jr.**[2], **Xiaoling Song**[3], **Johanna W. Lampe**[3,*]

[1]Department of Biostatistics, University of Nebraska Medical Center, Omaha, Nebraska, USA

[2]Department of Anesthesiology and Pain Medicine, University of Washington, Seattle, Washington, USA

[3]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[4]Department of Epidemiology, University of Washington, Seattle, Washington, USA

[5]Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, Wisconsin, USA

Abstract

Purpose: Objective biomarkers of dietary exposure are needed to establish reliable diet-disease associations. Unfortunately, robust biomarkers of macronutrient intakes are scarce. We aimed to assess the utility of serum, 24-hour urine and spot urine high-dimensional metabolites for the development of biomarkers of daily intake of total energy, protein, carbohydrate and fat, and the percent of energy from these macronutrients (%E).

[*]Corresponding author: jlampe@fredhutch.org.

Authors' contributions

The authors' responsibilities were as follows-CZ, DR, MLN, LFT, RLP, SAAB, and JWL: designed the research; MLN, LFT, RLP, JWL: conducted the feeding study; GANG, DR, DD, HG, RP, LB, GAB, XS: collected metabolite and biospecimen data; CZ and YZ: performed statistical analysis; CZ, GANG, DR, JWL: led the manuscript drafting; MLN, LFT, SAAB and RLP: provided critical review; JWL: had primary responsibility for the final content; and all authors: read and approved the final manuscript.

Declarations

Conflicts of interest/Competing interests

None of the authors reported a conflict of interest related to the study.

Ethics approval

The study was approved by the IRB of Fred Hutchinson Cancer Research Center.

Consent to participate

Consent to participate in research has been obtained.

Consent to publication

Consent to publication has been obtained.

Availability of data and material

Data used in this study will not be publicly available. Formal request via the Women's Health Initiative is needed to get access the data.

Code availability

R code used in this study will not be publicly available. Please request code access via email.

Methods: A 2-week controlled feeding study mimicking the participants' habitual diets was conducted among 153 postmenopausal women from the Women's Health Initiative (WHI). Fasting serum metabolomic profiles were analyzed using a targeted liquid chromatography-tandem mass spectrometry (LC-MS/MS) assay for aqueous metabolites and a direct injection based quantitative lipidomics platform. Urinary metabolites were analyzed using ^{1}H nuclear magnetic resonance (NMR) spectroscopy at 800 MHz and by untargeted gas chromatography-mass spectrometry (GC-MS). Variable selection was performed to build prediction models for each dietary variable.

Results: The highest cross-validated multiple correlation coefficients (CV-R^2) for protein intake (%E) and carbohydrate intake (%E) using metabolites only were 36.3% and 37.1% respectively. With the addition of established dietary biomarkers (doubly labeled water for energy and urinary nitrogen for protein), the CV-R^2 reached 55.5% for energy (kcal/d), 52.0% and 45.0% for protein (g/d, %E), 55.9% and 37.0% for carbohydrate (g/d, %E).

Conclusion: Selected panels of serum and urine metabolites, without the inclusion of doubly labeled water and urinary nitrogen biomarkers, gives a reliable and robust prediction of daily intake of energy from protein and carbohydrate.

Keywords

protein; carbohydrate; controlled feeding study; dietary biomarker; metabolomics; postmenopausal women

INTRODUCTION

It is important to understand how diet may alter the risk of chronic diseases, such as cardiovascular diseases (CVDs), diabetes, and cancer. Although there is evidence of associations between dietary patterns and certain chronic diseases [1], information on the relationships between intake of specific nutrients, foods, or dietary patterns and risk of chronic diseases is still limited by methodologic issues. One crucial obstacle is the reliance on self-report for dietary intake assessment, since self-report of macronutrient intake may incorporate substantial systematic error. Strong evidence has shown that the misreporting of dietary components is related to study participant characteristics (e.g., body mass index (BMI)) [2–4]. Such systematic measurement error can lead to bias that is not easily corrected [5]. Some previous work from the Women's Health Initiative (WHI) has demonstrated successful examples of approaches to correct the systematic errors using regression calibration methods for selected dietary components and their associations with the risks of certain major chronic diseases [6–11]. However, to correct for the systematic bias more generally, objectively measured biological indicators (i.e. biomarkers) of additional dietary features of interest are needed.

The macronutrients, carbohydrate, fat, and protein, comprise the principal sources of energy in the human diet. Urinary nitrogen (UN) has been identified as a robust biomarker of protein intake [12]. However, efforts to identify biomarkers for total carbohydrate and total fat intake have had limited success, with our recent work on plasma fatty acid profiles for carbohydrate as a possible exception [13]. We showed that serum phospholipid fatty acids (PLFA) combined with participant characteristics and established biomarkers could predict

consumed total carbohydrate (g/d), total saturated fatty acids (SFAs) (g/d), percentage of energy from SFAs (%E), and total trans fatty acids (g/d) with CV-R^2>36% [13]. It is likely impossible to use any single objectively measured metabolite to satisfactorily assess total carbohydrate and total fat intake because these nutrients are present in a variety of forms, and their metabolism is complex. For example, carbohydrate metabolites are central to many essential metabolic pathways [14], but these pathways are also interconnected [15]. Some ingested carbohydrates are metabolized through fatty acid pathways and can be stored as fat, which also makes total fat intake difficult to assess using biomarkers [13]. Identifying satisfactory objective measures of total carbohydrate and fat intake remains an important research gap.

The growing field of metabolomics promises unique opportunities for the identification of dietary biomarkers using biological specimens. It involves quantitative analysis of several hundreds to thousands of small molecules (MW ~1000 Da) in a variety of samples including blood and urine. Analysis of such metabolite data offers new avenues to objectively link metabolites with dietary patterns. Metabolite analysis most often utilizes two powerful analytical platforms, mass spectrometry (MS) [16] and nuclear magnetic resonance (NMR) spectroscopy [17]. There are many methods within the MS platform, such as liquid and/or gas chromatography resolved mass spectrometry (LC-MS or GC-MS) and lipidomics that are widely used in metabolomics studies. Each of these methods provides information on largely complementary sets of metabolites and hence the use of more than one analytical method for analysis of the same samples provides access to a wide and complementary pool of metabolites. Many metabolomics studies to date have focused on the identification of markers for a variety of dietary factors [18–20].

Food and nutrition play a critical role in human health and disease and interest to utilize metabolomics in this area is rapidly increasing [21]. In the present study, which is focused on the development of dietary biomarkers for carbohydrate, protein, and fat intake, we have performed metabolomics studies of serum and urine from a controlled feeding study from the WHI. Multiple analytical platforms of MS (LC-MS, direct injection MS, and GS-MS) and NMR spectroscopy were utilized to access the wide and complementary pools of metabolites, and to investigate new approaches to biomarker identification for dietary intake. In particular, comprehensive analysis of the data was made with the goal of evaluating the utility of the metabolomics data alone or in conjunction with participant characteristics and/or diet-related biomarkers such as estimated energy intake derived from doubly labeled water (DLW) and estimated protein intake using UN, for the development of biomarkers of carbohydrate, protein, and fat intake.

METHODS

Study Design

We carried out the current work using data and specimens from a recently completed controlled feeding study “Nutrition and Physical Activity Assessment Study-Feeding Study” (NPAAS-FS), an ancillary study to the WHI [22] which was designed to develop and evaluate potential intake biomarkers for nutrients and dietary components more generally. It was conducted in a subset of 153 women who were currently enrolled in a WHI

Extension Study from the Seattle WA area and who had been in the Observational Study cohort, the Dietary Modification trial Comparison group, or Hormone Therapy Trials. There are participant overlaps between this subgroup and two previously conducted studies: the Nutrition Biomarker Study (n=544 total; n=2 overlap) and the first cycle of NPAAS (n=450 total; n=14 overlap), which was observational in nature. The full inclusion criteria for the NPAAS-FS have been published [22]. We used a novel study design to conduct a controlled feeding study. Rather than feed all women the same standard diets, we designed an individual 2-week diet for each woman that mimicked her habitual diet as described by her 4-day food record (4FDR) with adjustment based on individual discussion with the study dietitian and energy adjustment based on calibrated estimates of total energy intake. This was provided as a 4-day rotating food plan with foods prepared by the Fred Hutchinson Cancer Research Center Human Nutrition Laboratory and provided to each participant. The goal of this approach was not to replicate precisely the diets of the study participants but to approximate them to minimize perturbations to blood and urine measures over the 2-week controlled feeding period, and to substantially preserve the normal variation in nutrient and food consumption in the study population. The details of the controlled feeding study including participant characteristics, and specimen collection procedures have been presented [22]. Participants were asked to eat the provided diets during the 14-day period and to complete a daily menu checklist to document deviations from the diet provided to each individual. Unconsumed food was returned to the study center, weighed and recorded. Nutrient content of the consumed diets was assessed using Nutrition Data System for Research software (version 2010; Nutrition Coordinating Center, University of Minnesota); these were the dietary variables used for model building. 24-hour and spot urine collections were made on day 13 of the 2-week feeding period and blood for serum was collected after a 12-hour overnight fast on day 14. DLW (in 2 spot urines on days 1 and 14) and UN (in the 24-hour urine) were measured according to established protocols [22]. Participant characteristics including dietary supplement use, season of participation, age, BMI and self-reported physical activity were collected at the time of enrollment in the NPAAS-FS, and other characteristics including race/ethnicity and education were measured at the time of enrollment in WHI.

Metabolite Profiling

Serum metabolite measurements—Serum samples from 153 NPAAS-FS participants, along with 17 split-sample blinded duplicates were analyzed by targeted LC-MS/MS using a Sciex Triple Quad 6500+ mass spectrometer. Specifically, serum samples were prepared by aqueous extraction of metabolites using methanol, as described previously [23]. The LC system was composed of four Shimadzu Nexera LC-20 pumps, an AB Sciex/CTC autosampler and AB Sciex column compartment containing a column-switching valve (AB Sciex, Toronto, ON, Canada). Two HILIC (hydrophilic interaction chromatography) columns (Waters XBridge Amide; 150 × 2.1 mm, 2.5 μm particle size), connected in parallel were used for positive and negative ionization modes. For chromatography, the mobile phase was composed of Solvent A: 10 mM ammonium acetate in 95% H_2O/3% acetonitrile/2% methanol/0.2% acetic acid, and Solvent B: 10 mM ammonium acetate in 93% acetonitrile/2% methanol/5% H_2O /0.2% acetic acid. Metabolites were analyzed in positive or negative ionization mode by injecting each sample twice, 5 μL for analysis using

positive ionization mode and 10 μL for analysis using negative ionization mode. The assay was developed using authentic commercially obtained compounds (Sigma-Aldrich, Saint Louis, MO or Fisher Scientific, Pittsburgh, PA) and targeted a total 303 metabolites that represented >40 different metabolic pathways, along with 33 stable-isotope labeled internal standards (Cambridge Isotope Laboratory, Tewksbury, MA). Metabolite concentrations were obtained using MultiQuant 3.0.2 software. A total of 303 metabolites were targeted, of which 155 were detected with less than 20% missing values.

Separately, quantitative analysis of serum lipids was performed using the Lipidyzer™ platform consisting of Shimadzu Nexera X2 LC-30AD pumps, a Shimadzu Nexera X2 SIL-30AC autosampler, and an AB Sciex Q TRAP® 5500 mass spectrometer equipped with SelexION® for differential mobility spectrometry (DMS) [24–25]. The method used 1-propanol as the chemical modifier for the DMS. Lipid metabolites were first extracted using dichloromethane/methanol to isolate the lipid species and remove proteins [24]. Samples were then introduced to the mass spectrometer by flow injection analysis at 8 uL/min. Each sample was injected twice, once with the DMS on and once with the DMS off. The method targeted 1070 different lipids that represented 13 different classes as follows: the DMS on mode detected phosphatidylcholine (PC), phosphatidylethanolamine (PE), lysophosphatidylcholine (LPC), lysophosphatidylethanolamine (LPE) and sphingomyelin (SM) lipid classes, whereas the DMS off mode detected cholesterol ester (CE), ceramide (CER), diacylglycerol (DAG), dihydroceramide (DCER), free fatty acid (FFA), hexosylceramide (HCER), lactosylceramide (LCER) and triacylglycerol (TAG) lipid classes. Data were acquired and processed using Analyst 1.6.3 and Lipidomics Workflow Manager 1.0.5.0. Absolute concentrations (in μM) were obtained based on a mixture of 54 isotope labeled internal standards, which were added to each sample during sample preparation. This approach resulted in the measurement of 664 serum lipids that had less than 20% missing values.

Urine metabolite measurements—Metabolite profiles of both spot and 24-hour urine samples (including split-sample blinded duplicates) were analyzed by ^{1}H NMR spectroscopy using a Bruker Avance III 800 MHz NMR spectrometer. For NMR analysis, urine samples (300 μL each) were mixed with an equal volume of phosphate buffer (100 mM, pH =7.4) containing an internal standard, TSP (3-(trimethylsilyl) propionic-2,2,3,3-d_4 acid sodium salt, 25 μM) and transferred to 5 mm NMR tubes. The samples were analyzed using a Bruker Avance III 800 MHz NMR spectrometer equipped with a cryogenically cooled probe and Z-gradients suitable for inverse detection. One dimensional NMR experiments using the '*noesypr1d*' pulse sequence with water suppression using presaturation, were performed using identical experimental conditions. Each spectrum was obtained using 10,000 Hz spectral width and 32,768 time-domain data points. Free induction decay (FID) signals were Fourier transformed after multiplying using an exponential window function and a line broadening of 0.5 Hz after setting the spectrum size to 32,768 points. Resulting spectra were phase and baseline corrected, and the chemical shifts were referenced to the internal TSP peak. Metabolites were then identified based on the literature and chemical shift databases [26–27]. Metabolite concentrations were obtained after normalizing NMR spectra with reference to the internal standard, TSP, peak. Bruker

Topspin versions 3.0 and 3.1 software packages were used for NMR data acquisition and processing, and Bruker AMIX software was used for metabolite quantitation. The NMR measurements were made in two batches at different times that spanned roughly a year. Relative concentrations for 57 metabolites were obtained. None of the metabolites had any missing values.

Separately, urine samples were analyzed using global GC-MS using an Agilent 7890/5975C GC−MS instrument and following established protocols [28]. Urine samples were treated with urease enzyme to deplete the urea level followed by methoxymation using methoxime. Urine metabolites were then derivatized using MSTFA (N-Methyl-N-(trimethylsilyl) trifluoroacetamide) with 1% (v/v) TMSC (trimethylchlorosilane) to facilitate volatilization of metabolites, which is required for GC-based analysis. Prior to derivatization, the samples were mixed with myristic acid-d_{27} and a FAME (fatty acid methyl-ester) mixture of retention time index compounds. The samples (1 μL) were injected onto the instrument using splitless mode. Helium was used as the carrier gas with a flow rate 1.2 mL/min. Separation was performed using an Agilent DB-5ms + 10 m Duraguard capillary column (30 m × 250 μm × 0.25 μm). The column temperature was maintained at 60 °C for one min, then increased at a rate of 10 °C/min to 325 °C and held at this temperature for 10 min. Mass spectral signals were collected after a solvent delay of 4.90 min. Peak intensities and elution times for the retention time index compounds were verified by m/z values after each experiment. After converting the data to the appropriate format, MS peaks were analyzed using Agilent MassHunter Quantitative Analysis software and PARADISe version 1.1.6 [29]. Relative concentrations of metabolites were obtained after normalizing the data with respect to the internal standard, myristic acid-d_{27}. This approach resulted in the identification of 285 metabolites, 275 for the 24-hour urine samples and 262 for spot urine that had less than 20% missing values.

Quality controls (QC) used in the metabolite analysis—Analysis protocols used multiple layers of QC samples as well as isotope labeled or unlabeled internal standards to assess instrument stability/performance during the analysis and help with normalization and metabolite quantitation. Different types of QCs used included: (a) unblinded instrument QC samples (commercially obtained pooled human serum from Innovative Research, Inc. (Novi, MI)) run every 10 samples and at the beginning and end of each batch of samples; (b) blinded, pooled study samples (5% for urine; 10% for serum) interspersed with the biological study samples (3 QCs/batch of 27 study samples), used to normalize batches of samples over the run; (c) 17 split-sample blinded duplicates of study samples also interspersed with study serum and urine samples, used to calculate reported average metabolite CV values; (d) isotope labeled internal standards for targeted analysis of aqueous metabolites (n=33) and lipids (n=54) in serum, which enabled absolute concentration determination and ensured evaluation of instrument stability and data quality; (e) internal standard, TSP, used to assess the spectral quality, calibrate spectra, and help with data normalization of urine NMR spectra; and (f) FAME (fatty acid methyl esters) of different fatty acid chain length and myristic acid-d_{27} retention time index compounds for help with metabolite identification and data normalization. Table 2 shows a summary list of metabolites detected in serum and urine using the different analytical platforms. Median

CVs of blinded pooled study QC samples/batch for the 4 different platforms (two for serum analysis and 2 for urine analysis) across the samples were 4.0%/1.2% for global NMR from 24-hour/spot urine, 5.5% for targeted lipidomics, 7.2% for targeted LC-MS/MS and 31.3% for global GC-MS platforms.

Serum phospholipid fatty acid measurements—PLFA were measured by the Public Health Sciences Biomarker Lab at the Fred Hutchinson Cancer Research Center using GC and the Folch extraction method [30] as described [13]. Phospholipids were separated from other lipids by one dimensional thin layer chromatography [31]. Fatty acid methyl esters of the phospholipids were prepared by direct transesterification [32] and separated using GC (Agilent 7890 Gas Chromatograph with flame ionization detector, Supelco fused silica 100 m capillary column SP-2560). The relative concentration of each fatty acid was expressed as a weight percentage of total PLFA analyzed (i.e., the sum of the 41 FA was 100%). A lab QC sample (pooled plasma) was included with each batch of study samples. Inter-batch CV for the lab QC sample was <12.7% (median 2.6%) for all PLFA except for one very minor fatty acid 20:3n3 (<0.1% by weight percentage), which had CV of 27.1%.

Statistical Analyses

Data preprocessing—For metabolomics variables, those with more than 20% missing values were removed to ensure robust results. For the remaining variables, half of the minimum positive value was used to impute the values that were below detection limits. Because some participants' body weights had slight changes during the 2-week period (weight change ranged from −3.6 to 2.4 kg), we used weight change and DLW-based total energy expenditure (TEE) to calculate the biomarker for total energy intake (Ein). Precise dietary intake data used as the outcome variable were derived from nutrient analysis of the consumed foods during the feeding study as described above (Study Design). Dietary and all lab-measured variables were log-transformed to be consistent with other analyses in the NPAAS-FS [22]. Outliers were truncated to Q1–3*IQR or Q3+3*IQR [33] where IQR represents the interquartile range and Q1 and Q3 represent the first and the third quartiles. Normalization was performed for LC-MS and GC-MS data using local polynomial regression fitting (loess) over run order with the span parameter set at 0.75 within each batch among QC samples. Given the high correlation between concentration and composition measures for serum samples, we chose composition measures for the analysis of this paper given their smaller CV for lipidomic data and larger number of features identified for LC-MS data. For ease of interpretation, un-identified metabolites from GC-MS were removed from the analysis.

Model Building—The model building involved two steps. First, we ran a LASSO [34] model for the regression of each macronutrient intake variable (derived from the consumed menus) on all the metabolites. A penalty parameter used to limit the number of variables included in the model was selected by 5-fold cross-validation [35] (with the restriction that the maximum number of variables selected be less than 15). The prediction model was built with the second round of linear regression after variable selection [36]. To account for the variation in variable selection and estimate the cross-validated percent of variation (CV-R^2) explained by the regression model, we adopted the refitted approach [37] by estimating the

cross-validated residual variance. We performed the following procedure 100 times and took the average. For each run, we randomly split the data into two sets with roughly equal size. Within each data set, we ran variable selection as described above and evaluated the residual variance from the other data set. We considered the threshold of CV-R^2 >36% as a standard for a useful biomarker [22]. Separate models were developed when only serum was used, only urine was used, and both were used. Separate models were built for macronutrient intake (g/d) and %E. We also performed the same analysis but replaced the 24-hour urine metabolite variables with spot urine variables and adjusted for log-transformed creatinine.

To evaluate the utility of existing biomarkers (Ein and UN) and participant characteristics in terms of increasing prediction accuracy for macronutrient intake, we ran a sequence of models using the general approach as indicated above but with a varying list of variables to be selected: Model 1 used metabolomic data only. Model 2 used metabolomic data and participant characteristics (including dietary supplement use, race/ethnicity, season of participation, education, age, BMI and self-reported physical activity). Model 3 used metabolomic data, participant characteristics, and diet-related biomarkers (Ein and UN). For each dietary outcome of interest, prediction equations from those models with the largest CV- R^2 that passed the 36% CV-R^2 threshold were calculated. The CV-R^2 values for each specific variable were computed as R^2 for that specific variable multiplied by the CV-R^2 for the whole model and then divided by R^2 for the whole model. As the variables included in our model were correlated, the order of variables entered into the model affected the R^2 for each specific variable; thus we chose to calculate the R^2 and CV-R^2 based on the decreasing order of the absolute value of each variable's standardized regression coefficients. Additional models were applied to evaluate whether the prediction performance could be improved by adding plasma PLFAs to the list of variables to be selected. Since a previous study showed the potential utility of self-reported diet intake data [38], we also explored how adding baseline WHI FFQ data collected at WHI cohort entry might improve the model. Given the measurement error in the data from the FFQ collected at the time of the feeding study might be highly correlated with that of the 4DFR completed right before the feeding study, we chose instead to use the "baseline" FFQ data. Here "baseline" is considered as year 1 for women in the WHI Dietary Modification Trial (WHI-DM) and year 0 for the others; that is, the time of cohort entry. All statistical analyses were performed using R4.0.2.

RESULTS

The individual characteristics of the 153 participants of the NPAAS-FS are reported in Table 1. Most of the women were white, with substantial education (some college or associate degree or higher) and used dietary supplements. The sample showed a good range of macronutrient intakes, with the upper end of the 95% range approximately double that of the lower end of the range. Table 2 shows the number of measured metabolites from the metabolomics platforms and the stable features that were available in ≥80% of the samples. From the coefficients of variation, we can see that the most precise signals came from NMR followed by the lipidomics and targeted LC-MS. The QC data suggest that GC-MS had the highest coefficients of variation.

Table 3 shows the prediction accuracy (CV-R^2) for total energy intake (kcal/d), and intakes (g/d and %E) of total protein, total carbohydrate, and total fat when using metabolites with and without other information including participant characteristics and diet-related biomarkers (Ein and UN). Without participant characteristics, the metabolites themselves predicted protein intake (%E) and carbohydrate intake (%E) adequately with CV-R^2 of 36.3% and 37.1% respectively. Adding participant characteristics did not show meaningful improvement. Adding the diet-related biomarkers improved the prediction of total carbohydrate intake (g/d) from a CV-R^2 of 36.1% to 57.0%. Adding the PLFAs did not further improve the carbohydrate predictions. Adding the diet-related biomarkers also improved the prediction of total protein intake (g/d) from a CV-R^2 27.4% to 52.0%. The metabolites predicted fat intake (g/d) poorly with the largest CV-R^2 at 3.5% without diet-related biomarkers included. Adding the diet-related biomarkers and participant characteristics increased the CV-R^2 to only 21.0%. Inclusion of the baseline FFQ improved the CV-R^2 for protein (g/d) to 54.7% and protein (%E) to 48.0% (data not shown). No improvement was observed for carbohydrate or fat.

We also present the best prediction models (with the largest CV-R^2) for each dietary measure if it reached our ≥36% CV-R^2 criteria. Table 4 shows the best prediction model for protein (%E) without established diet-related biomarkers. Tables 5 and 6 show the best prediction model for carbohydrate (g/d) and (%E), respectively, without the inclusion of established diet-related biomarkers. Table S1–S5 in the supplemental material show the improved prediction models for energy (kcal/d), protein (g/d), protein (%E), carbohydrate (g/d), and carbohydrate (%E) with established diet-related biomarkers included. Models including both established biomarkers and the metabolite data improved the performance compared to using only diet-related biomarkers for energy and protein. Model checking did not show severe nonlinearity, outliers or high influential points for the final fitted models.

DISCUSSION

In this study of 153 postmenopausal women enrolled in the WHI, we evaluated the application of metabolomics data (i.e. metabolites) obtained from serum and urine plus participant characteristics for the generation of prediction equations of intake of total protein, carbohydrate, and fat intake (g/d and %E). Overall, our analysis suggests that using metabolites themselves can achieve a useful prediction of carbohydrate intake (g/d and %E) and protein intake (%E) in this sample of women. However, the prediction of energy intake (kcal/d) and protein intake (g/d) still requires the availability of the established diet-related biomarkers (Table S1–2) although metabolites can provide some additional prediction power. Also, the prediction of fat intake (g/d) might require the measurement of additional or different objective measures and/or the use of more complex nonlinear models. As with most dietary biomarkers measured in blood and urine, the metabolites that predicted macronutrient intakes in our study reflected not only intake, but the physiologic processes involved in the absorption, metabolism, and excretion of these dietary constituents. This was evidenced by the contribution of urinary nitrogenous compounds to prediction of protein intake and de novo produced lipids to carbohydrate intake.

We tested several prediction models to explain protein intake (g/d and %E) (Table 3). Evaluation of metabolites alone (Model 1) showed that 24.0% of the variation in protein intake (g/d) could be explained. Adding participant characteristics, namely BMI, improved the predication of the CV-R^2 (Model 2) to 27.4%. UN contributed an additional ~20% to the CV-R^2 (Model 3). For the prediction of %E from protein, we found that Model 2 provide a CV-R^2 of 36.6% and adding UN increased CV-R^2 to 45.0%. When UN is included, the prediction models for protein (g/d) suggests that the spot urine contains better information than 24-hour urine given its larger point estimate for CV-R^2, though the opposite was true for prediction for protein (%E). However, without UN, the prediction models using 24-hour urine outperform those using spot urine. This is not unexpected because when UN is included (Model 3), the spot urine model uses information from both 24-hour urine (for UN) and spot urine. However, the improvement with the addition of spot urine is limited (CV-R^2 change from 48.1% to 52.0%). The sequential addition of variables from Model 1 to Model 3 showed that UN was effective for the prediction for protein (g/d) but did not provide much improvement for protein (%E) for both 24 hour-urine and spot urine models. The metabolites contributing to the variation in %E from protein in Model 1 (Table 4), as well as Model 3 for g/d protein (Table S2), included urinary metabolites known to contribute to the measure of UN (e.g., urea, creatine, and several amino acid derivatives). Interestingly, urinary propanediol and LPE 16:0 contributed 10.1% and 9.8% to the CV-R^2 for protein (%E), respectively, but were not selected for the prediction model for protein (g/d). Our primary analysis focused on identified metabolites. We conducted a secondary analysis, which included unidentified metabolites in the list of variable selection; this did not show any difference in CV-R^2. Also, we ran a sensitivity analysis excluding participants with more than a 1 kg/week (i.e., 2 kg during the 14-day period) weight change; the CV-R^2 did not have a meaningful change (<2%).

We also tested several prediction models to explain carbohydrate intake (g/d and %E) (Table 3). Evaluation of metabolites alone (Model 1) showed that 36.1% of the variation in carbohydrate intake (g/d) could be explained. Adding participant characteristics did not improve the prediction (Model 2). Ein typically contributed at least an additional 20% to the CV-R^2 for each type of biologic sample (Model 3). For the prediction of %E from carbohydrate, we found that Model 1 provided a CV-R^2 37.1% and adding Ein did not lead to a meaningful change in CV-R^2. This suggests that an objective measure of estimated energy intake is an important contributor to predicting total carbohydrate intake (g/d) and is likely due to its correlation with total energy intake rather than specific information on carbohydrate consumption in this study sample, consistent with a primary role of carbohydrate in providing energy.

We observed that the carbohydrate models built on serum + spot urine and on serum + 24 hour urine metabolites (Table 5 and Table 6) contained very few urine metabolites; the serum metabolites contributed almost all additional information. This may be based on the fact that most of the metabolites related to carbohydrate and energy metabolism were detected in serum. With a focus on total carbohydrate from a variety of food sources rather than on specific carbohydrate-rich foods, small molecules detected in urine and often associated with particular foods (e.g., alkylresorcinols from certain whole grains) are probably less likely to contribute to a total carbohydrate model. Here, the metabolites

contributing to estimates of carbohydrate intake were predominantly TAGs and other lipid species, detected on the lipidomics platform. This finding was anticipated given that dietary carbohydrate provides much of the substrate acetyl-CoA used in de novo lipogenesis. Further, we showed previously in the NPAAS-FS that serum PLFA concentrations and participant characteristics explained 37.1% and 27.3% of the variation in total carbohydrate intake (g/d) and (%E), respectively [13]. Few aqueous metabolites contributed to the models; however, maltose, a disaccharide produced by the breakdown of starch, contributed 5.5% to the CV-R^2 carbohydrate (%E). These findings further support the potential utility of lipid biomarkers in the characterization of carbohydrate intake.

Though the metabolites in combination with diet-related biomarkers and some participant characteristics can successfully predict energy intake, protein intake and carbohydrate intake, we did not find a good biomarker for fat intake. For carbohydrate intake, adding the PLFAs as used in a previous study [13] into the list of variables to be selected did not improve the prediction, which suggests that the information of PLFAs likely overlaps with the metabolite data generated on the platforms considered here.

Our use of controlled feeding of women's habitual diets and the variety of food sources of carbohydrate within their diets provides a useful approach for biomarker development. The application illustrates how to characterize biomarkers (Models 1–3) using several metabolomics platforms for serum and urine. With the cross-validated LASSO, we were able to handle the high-dimensionality of the metabolomic data and produce a valid prediction model that was not over-fitted. However, the small sample size limited us to test the prediction accuracy between different models among these postmenopausal women as the confidence interval of the CV-R^2 tended to be wide. Also, the LASSO method assumes an additive effect of different metabolomics which might not be strictly true as the metabolomic pathways are complex. Metabolites within a pathway, or even across pathways can be correlated [39]. An alternative approach may be to sum together relevant metabolomics in the same pathway and then log transform the sum and include it as an additive term. The distribution in race/ethnicity and education level of the WHI study participants may limit generalization of the results to a broader population of postmenopausal women. Further, testing is needed of the transferability of the proposed biomarkers to other populations and the reproducibility of these metabolites measured in other labs using similar techniques.

There are a few limitations of the current study. First, the serum after overnight fasting and spot or 24-hour urine samples may not fully represent a full 14-day intervention. Longitudinal measurements of the metabolites might provide more insights in the relationships between metabolites and dietary responses; however, these are often the types of biospecimens available in cohorts, to which results from this study may be applied extending generalizability. Second, we measured UN in a single 24-hour urine which may be less reliable than multiple measurements [12]. Third, the prevalence of alcohol use was very limited among this population and thus we were not able to evaluate the impact of alcohol intake on these metabolite-based biomarkers which might prevent us from a direct generalization of our prediction equations to other populations. Finally, while the proposed novel biomarkers met a CV-R^2> 36% criterion that is motivated by R^2 values

of about 50% for the DLW energy intake assessment, and about 40% for the UN protein intake assessment, even larger CV-R^2 values for our proposed novel biomarkers would be preferable, and suitability also requires substantial sensitivity and specificity of the biomarker for the nutritional variable under consideration.

In conclusion, this analysis supports the utility of a metabolomics approach for the development of nutritional biomarkers of carbohydrate intake (g/d) and of relative intakes of protein and carbohydrate (%E). Thus, future cohort studies looking at associations between these variables and disease risk may benefit from measuring serum/urine samples to establish biomarker-based calibrated estimates of intake. For protein (g/d), the measure of UN was still essential to obtain a sufficiently robust biomarker although our analysis shows that using information from Ein and metabolites can further improve the prediction. For fat intake, the currently measured metabolites do not provide an adequate prediction, at least using linear models as we have done here. Further study might consider measurement of other metabolites and/or use of nonlinear predictions to find a better prediction model. Overall, this analysis also suggests that multi-platform metabolite-based biomarker profiles may warrant exploration for characterizing subtypes of protein, carbohydrate and fat, as well as various markers for food sources of these macronutrients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by National Cancer Institute grant R01 CA119171 and Office of Research Infrastructure Programs grant S10 OD021562. The Women's Health Initiative (WHI) is supported by the National Heart, Lung, and Blood Institute, NIH, US Department of Health and Human Services through contracts HHSN268201600046C (Fred Hutchinson Cancer Research Center), HHSN268201600001C (State University of New York, Buffalo), HHSN268201600002C (The Ohio State University), HHSN268201600003C (Stanford University), HHSN268201600004C (Wake Forest University), and HHSN271201600004C (WHI Memory Study) and grants P30 CA015704 and P30 CA023074.

The authors acknowledge the following investigators in the WHI Program:

Program Office: (National Heart, Lung, and Blood Institute, Bethesda, Maryland) Jacques Rossouw, Shari Ludlam, Joan McGowan, Leslie Ford, and Nancy Geller

Clinical Coordinating Center: (Fred Hutchinson Cancer Research Center, Seattle, WA) Garnet Anderson, Ross Prentice, Andrea LaCroix, and Charles Kooperberg

Investigators and Academic Centers: (Brigham and Women's Hospital, Harvard Medical School, Boston, MA) JoAnn E. Manson; (MedStar Health Research Institute/Howard University, Washington, DC) Barbara V. Howard; (Stanford Prevention Research Center, Stanford, CA) Marcia L. Stefanick; (The Ohio State University, Columbus, OH) Rebecca Jackson; (University of Arizona, Tucson/Phoenix, AZ) Cynthia A. Thomson; (University at Buffalo, Buffalo, NY) Jean Wactawski-Wende; (University of Florida, Gainesville/Jacksonville, FL) Marian Limacher; (University of Iowa, Iowa City/Davenport, IA) Jennifer Robinson; (University of Pittsburgh, Pittsburgh, PA) Lewis Kuller; (Wake Forest University School of Medicine, Winston-Salem, NC) Sally Shumaker; (University of Nevada, Reno, NV) Robert Brunner

Women's Health Initiative Memory Study: (Wake Forest University School of Medicine, Winston-Salem, NC) Mark Espeland

For a list of all the investigators who have contributed to WHI science, please visit: https://www-whi-org.s3.us-west-2.amazonaws.com/wp-content/uploads/WHI-Investigator-Long-List.pdf

REFERENCES

1. Jeppesen J, Schaaf P, Jones C, Zhou MY, Chen YD, Reaven GM (1997) Effects of low-fat, high-carbohydrate diets on risk factors for ischemic heart disease in postmenopausal women. Am J Clin Nutr 65:1027–1033. [PubMed: 9094889]
2. Neuhouser ML, Tinker L, Shaw PA, Schoeller D, Bingham SA, Horn LV, Beresford SA, Caan B, Thomson C, Satterfield S, Kuller L, Heiss G, Smit E, Sarto G, Ockene J, Stefanick ML, Assaf A, Runswick S, Prentice RL (2008) Use of recovery biomarkers to calibrate nutrient consumption self-reports in the Women's Health Initiative. Am J Epidemiol 167:1247–1259. [PubMed: 18344516]
3. Prentice RL, Willett WC, Greenwald P, Alberts D, Bernstein L, Boyd NF, Byers T, Clinton SK, Fraser G, Freedman L, Hunter D, Kipnis V, Kolonel LN, Kristal BS, Kristal A, Lampe JW, McTiernan A, Milner J, Patterson RE, Potter JD, Riboli E, Schatzkin A, Yates A, Yetley E (2004) Nutrition and physical activity and chronic disease prevention: research strategies and recommendations. J Natl Cancer Inst 96:1276–1287. [PubMed: 15339966]
4. Prentice RL, Mossavar-Rahmani Y, Huang Y, Van Horn L, Beresford SA, Caan B, Tinker L, Schoeller D, Bingham S, Eaton CB, Thomson C, Johnson KC, Ockene J, Sarto G, Heiss G, Neuhouser ML (2011) Evaluation and comparison of food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. Am J Epidemiol 174:591–603. [PubMed: 21765003]
5. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective. CRC Press, New York.
6. Zheng C, Beresford SAA, Van Horn L, Tinker LF, Thomson CA, Neuhouser ML, Di C, Manson JE, Mossavar-Rahmani Y, Seguin R, Manini T, LaCroix AZ, Prentice RL (2014) Simultaneous association of total energy consumption and activity-related energy expenditure with cardiovascular disease, cancer, and diabetes risk among postmenopausal women. Am J Epidemiol 180:526–535. [PubMed: 25016533]
7. Beasley JM, LaCroix AZ, Larson J, Huang Y, Neuhouser ML, Tinker LF, Jackson RD, Snetselaar L, Johnson K, Eaton C, Prentice RL (2014) Biomarker-calibrated protein intake and bone health in the Women's Health Initiative clinical trials and observational study. Am J Clin Nutr 99:934–940. [PubMed: 24552750]
8. Huang Y, Van Horn L, Tinker LF, Neuhouser ML, Carbone L, Mossavar-Rahmani Y, Thomas F, Prentice RL (2013) Measurement error corrected sodium and potassium intake estimation using 24-hour urinary excretion. Hypertension 63:238–244. [PubMed: 24277763]
9. Prentice RL, Neuhouser ML, Tinker LF, Pettinger M, Thomson CA, Mossavar-Rahmani Y, Thomas F, Qi L, Huang Y (2013) An exploratory study of respiratory quotient calibration and association with postmenopausal breast cancer. Cancer Epidemiol Biomarker Prev 22:2374–2383.
10. Beasley JM, Wertheim BC, LaCroix AZ, Prentice RL, Neuhouser ML, Tinker LF, Kritchevsky S, Shikany JM, Eaton C, Chen Z, Thomson CA (2013) Biomarker-calibrated protein intake and physical function in the Women's Health Initiative. J Am Gerontol Soc 61:1863–1867.
11. Neuhouser ML, Di C, Tinker LF, Thomson C, Sternfeld B, Mossavar-Rahmani Y, Stefanick ML, Sims S, Curb JD, LaMonte M, Seguin R, Johnson KC, Prentice RL (2013) Physical activity assessment: biomarkers and self-report of activity-related energy expenditure in the WHI. Am J Epidemiol 177:576–585. [PubMed: 23436896]
12. Bingham SA (2003) Urine nitrogen as a biomarker for the validation of dietary protein intake. J Nutr 133:921S–924S. [PubMed: 12612177]
13. Song X, Huang Y, Neuhouser ML, Tinker LF, Vitolins MZ, Prentice RL, Lampe JW (2017) Dietary long-chain fatty acids and carbohydrate biomarker evaluation in a controlled feeding study in participants from the Women's Health Initiative cohort. Am J Clin Nutr 105:1272–1282. [PubMed: 28446501]
14. Da Poian AT, Bacha T, Luz MRMP (2010) Nutrient utilization in humans: metabolism pathways. Nature Education 3:11

15. Minehira K, Bettschart V, Vidal H, Vega N, Di Vetta V, Rey V, Schneiter P, Tappy L (2003) Effect of carbohydrate overfeeding on whole body and adipose tissue metabolism in humans. Obes Res 11:1096–103. [PubMed: 12972680]
16. Raftery D (ed.) (2014) Mass spectrometry in metabolomics: methods and protocols. Methods in Molecular Biology, Vol. 1198. Humana Press/Springer Science, New York.
17. Nagana Gowda GA, Raftery D (eds.) (2019) NMR based metabolomics: methods and protocols. Methods in Molecular Biology, Vol. 2037. Humana Press/Springer Science, New York.
18. Clarke ED, Rollo ME, Pezdirc K, Collins CE, Haslam RL. (2020) Urinary biomarkers of dietary intake: a review, Nutr Rev. 78(5):364–381. [PubMed: 31670796]
19. Guasch-Ferré M, Bhupathiraju SN, Hu FB. (2018) Use of Metabolomics in Improving Assessment of Dietary Intake. Clin Chem. 64(1):82–98. [PubMed: 29038146]
20. Gibbons H, Brennan L. (2017) Metabolomics as a tool in the identification of dietary biomarkers, Proc Nutr Soc. 76(1):42–53. [PubMed: 27221515]
21. Nagana Gowda GA, Alvarado LZ, Raftery D (2017) Nutrition in the prevention and treatment of disease, 4th edn. Elsevier Inc, New York, pp.103–122.
22. Lampe JW, Huang Y, Neuhouser ML, Tinker LF, Song X, Schoeller DA, Kim S, Raftery D, Di C, Zheng C, Schwarz Y, Van Horn L, Thomson CA, Mossavar-Rahmani Y, Beresford SAA, Prentice RL (2017) Dietary biomarker evaluation in a controlled feeding study in women from the women's health initiative cohort. Am J Clin Nutri 105:466–475.
23. Navarro SL, Tarkhan A, Shojaie A, Randolph TW, Gu H, Djukovic D, Osterbauer KJ, Hullar MA, Kratz M, Neuhouser ML, Lampe PD, Raftery D, Lampe JW (2019) Plasma metabolomics profiles suggest beneficial effects of a low–glycemic load dietary pattern on inflammation and energy metabolism. Am J Clin Nutr 110:984–992. [PubMed: 31432072]
24. Hanson AJ, Banks WA, Bettcher LF, Pepin R, Raftery D, Craft S (2020) Cerebrospinal fluid lipidomics: effects of an intravenous triglyceride infusion and apoE status. Metabolomics 16:6.
25. Dibay Moghadam S, Navarro SL, Shojaie A, Randolph TW, Bettcher LF, Le CB, Hullar MA, Kratz M, Neuhouser ML, Lampe PD, Raftery D, Lampe JW (2020) Plasma lipidomic profiles after a low and high glycemic load dietary pattern in a randomized controlled crossover feeding study. Metabolomics 16:121. [PubMed: 33219392]
26. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djoumbou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorndahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A (2013) HMDB 3.0--The human metabolome database in 2013. Nucleic Acids Res 41:D801–807. [PubMed: 23161693]
27. Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Kent Wenger R, Yao H, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–408. [PubMed: 17984079]
28. Chan ECY, Pasikanti KK, Nicholson JK (2011) Global urinary metabolic profiling procedures using gas chromatography-mass spectrometry. Nature protocols 6:1483–1499. [PubMed: 21959233]
29. Johnsen LG, Skou PB, Khakimov B, Bro R (2017) Gas chromatography - mass spectrometry data processing made easy. Journal of Chromatography A 1503:57–64. [PubMed: 28499599]
30. Folch J, Lees M, Sloane Stanley GH (1957) A simple method for the isolation and purification of total lipides from animal tissues. J Biol Chem 226:497–509. [PubMed: 13428781]
31. Schlierf G, Wood P (1965) Quantitative determination of plasma free fatty acids and triglycerides by thin-layer chromatography. J Lipid Res 6:317–319. [PubMed: 14328439]
32. Lepage G, Roy CC (1986) Direct transesterification of all classes of lipids in a one-step reaction. J Lipid Res 27:114–120. [PubMed: 3958609]
33. Hubert M, Van der Veeken S (2007) Outlier detection for skewed data. Journal of Chemometrics 22:235–246.
34. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58:267–288.
35. Kohavi RA (1995) Study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. San Mateo, CA: Morgan Kaufmann. 2:1137–1143.

36. Belloni A, Chernozhukov V (2013) Least squares after model selection in high-dimensional sparse models. Bernoulli 19:521–547.

37. Fan J, Guo S, Hao N (2012) Variance estimation using refitted cross-validation in ultrahigh dimensional regression. J R Stat Soc Series B Stat Methodol 74:37–65. [PubMed: 22312234]

38. Prentice RL, Pettinger M, Neuhouser ML, Tinker LF, Huang Y, Zheng C, Manson JE, Mossavar-Rahmani Y, Anderson GL, Lampe JW (2020) Can dietary self-reports usefully complement blood concentrations for estimation of micronutrient intake and chronic disease associations? Am J Clin Nutr 112:168–179. [PubMed: 32133498]

39. Rosato A, Tenori L, Cascante M, De Atauri Carulla PR, Martins dos Santos VAP, Saccenti E (2018) From correlation to causation: analysis of metabolomics data using systems biology approaches. Metabolomics 14:37. [PubMed: 29503602]

Table 1.

Participant characteristics and nutrient intakes of women in the Women's Health Initiative Nutrition and Physical Activity Assessment Feeding Study (n=153)

Variable (category)	**N**	**%**
Age (year)		
60–69	10	7.0
70–79	127	83.0
80–85	16	10.0
Race/Ethnicity		
Caucasian	146	95.4
Non-Caucasian	7	4.6
BMI (kg/m^2)		
Normal (<25.0)	61	39.9
Overweight (25–30)	60	39.2
Obese (≥30)	32	20.9
Use of Any Dietary Supplement	130	85.0
Current smoking	3	2.0
Season of study participation		
Spring	38	24.8
Summer	51	33.3
Fall	31	20.3
Winter	33	21.6
Years of education		
High school/General Educational Development diploma	10	6.5
Schooling after high school	60	39.2
College degree or higher	82	53.6
Missing	1	0.7
Recreational physical activity (MET/week)		
0–5.5	39	25.5
5.6–12.25	38	24.8
12.3–24.0	39	25.5
>24	37	24.2
Nutrient intake *	Geometric Mean	95% Range
Energy (kcal/d)	1904	1417, 2552
Protein (g/d)	78	50, 113
Carbohydrate (g/d)	212	130, 331
Fat (g/d)	80	51, 130
Protein (%E)	16.3	11.6, 22.1
Carbohydrate (%E)	44.6	28.4, 56.3

Variable (category)	N	%
Fat (%E)	37.6	26.9, 49.9
Total Dietary Fiber (g/d)	24.5	14.6, 39.8
Total dietary folate equivalents (μg/d)	948.5	317.4, 3335.4
Vitamin B-12 (μg/d)	25.4	2.2, 1860.8
Vitamin C(mg/d)	238.2	49.8, 1436.2

* Based on diet as consumed during the feeding study.

Table 2.

Number of metabolites identified in each platform and their median coefficients of variation (CV %) across the specimens from the 153 women in the controlled feeding study.

Platform and biologic sample	Features (n)		CV[*] (%)
	Total[1]	<20% Missing	
LC-Q-TOFMS serum (Composition)	1070	664	5.5
Targeted LC-MS serum (Composition)	303	155	7.2
GC-MS 24-hour urine	285	275[#]	31.3
GC-MS spot urine	285	262[&]	31.3
NMR 24-hour urine	57	57	4.0[a]
NMR spot urine	57	57	1.2[a]

[1] Total number of identified metabolites/features

[#] 137 features are un-identified

[&] 127 features are un-identified

[*] CV among those features with <20% missing.

[a] NMR measurements were made in two batches spaced in time by ~1 year.

Table 3.

Cross-validated R^2 using metabolite data alone on predicting dietary intakes of energy, protein, carbohydrate and fat.

	Energy (kcal/d)									
	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine
Model 1[a]	12.8%	1.4%	12.0%	0.4%	10.1%					
Model 2[b]	11.6%	0.8%	11.2%	0.7%	10.7%					
Model 3[c]	55.1%	53.3%	55.0%	53.8%	55.5%					
	Protein (g/d)					Protein (%E)				
	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine
Model 1	23.8%	17.1%	24.0%	5.5%	20.4%	30.5%	36.0%	36.3%	19.8%	28.3%
Model 2	27.4%	24.8%	26.5%	7.5%	24.9%	28.8%	36.6%	33.9%	18.4%	27.0%
Model 3	48.1%	47.8%	47.7%	49.3%	52.0%	34.9%	45.0%	40.9%	32.8%	34.8%
	Carbohydrate (g/d)					Carbohydrate (%E)				
	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine
Model 1	34.2%	6.3%	33.5%	1.4%	36.1%	32.8%	22.8%	37.1%	2.1%	30.4%
Model 2	33.7%	5.2%	33.0%	1.2%	35.7%	32.2%	22.6%	36.2%	2.4%	29.9%
Model 3	55.6%	40.4%	55.9%	34.2%	57.0%	33.8%	26.5%	37.0%	5.4%	33.0%
	Fat (g/d)					Fat (%E)				
	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine	Serum	24-hour Urine	Serum + 24-hour Urine	Spot Urine	Serum + Spot Urine
Model 1	2.4%	0.7%	2.4%	1.0%	3.5%	12.6%	0.9%	11.1%	1.6%	13.9%
Model 2	1.8%	0.4%	1.5%	1.1%	2.3%	12.8%	1.2%	11.8%	1.8%	14.4%
Model 3	21.0%	19.9%	20.4%	19.2%	20.9%	10.3%	1.2%	9.1%	1.5%	10.1%

[a]Model 1: Metabolites only

[b]Model 2: Metabolites + Participant Characteristics

[c]Model 3: Metabolites + Participant Characteristics + Diet-related Biomarker (Total Energy Intake, Urine Nitrogen)

[*]Serum metabolites were measured by LC-MS/MS and direct-injection MS/MS (lipidomics). 24-hour and spot urinary metabolites were measured by NMR and GC-MS. See text for additional information.

Table 4.

Variables selected for predicting dietary protein intake (E%) with their R^2 and corresponding cross-validated R^2 (%) using serum + 24-hour urine metabolites without participant characteristics or established diet-related biomarkers (Model 1).

Variable[#]	Coefficient	R^2	CV-R^2
Intercept	−2.836		
Propanediol (urine)	−0.082	17.6%	10.1%
Lysophosphatidylethanolamine (LPE 16:0[*]) (serum)	0.251	17.1%	9.8%
Urea (urine)	0.138	5.7%	3.3%
Arabitol/Xylitol (serum)	−0.101	5.3%	3.1%
Creatine (serum)	0.065	4.8%	2.8%
2-Oxoisovalerate (serum)	0.125	4.2%	2.4%
2-Hydroxyglutarate (serum)	−0.136	3.4%	1.9%
Maltose (urine)	−0.019	1.7%	1.0%
Methyl-glycocholate (urine)	−0.046	1.4%	0.8%
2-Hydroxybutyrate (serum)	0.057	0.8%	0.5%
Cholesteryl ester (CE 18:3[*]) (serum)	−0.075	0.6%	0.4%
Glutamine (serum)	−0.111	0.3%	0.2%
Uridine (serum)	0.052	0.2%	0.1%
Cholesteryl ester (CE 22:6[*]) (serum)	0.020	0.1%	0.0%
Creatine (urine)	−0.001	0.0%	0.0%
Total		63.2%	36.3%

[#]Reordered by R^2

[*]In LPE 16:0, 16 indicates number of carbons and 0 indicates number of double bonds in the fatty acid chain.

LPE: lysophosphatidylethanolamine. CE: cholesterol ester.

Table 5.

Variables selected for predicting dietary carbohydrate intake (g/d) with their R^2 and corresponding cross-validated R^2 (%) using serum + spot urine metabolites without participant characteristics or established diet-related biomarkers (Model 1).

Variable[#]	Coefficient	R^2	CV-R^2
Intercept	9.770		
Triacylglycerol (TAG 50:5(FA18:2)[*]) (serum)	0.295	25.0%	14.5%
Lysophosphatidylcholine (LPC 22:5[*]) (serum)	0.102	8.6%	5.0%
Triacylglycerol (TAG 50:4(FA18:1)[*]) (serum)	−0.332	7.6%	4.4%
Cholesteryl ester (CE 22:2[*]) (serum)	0.221	5.9%	3.4%
Indole-3-propionate (serum)	0.084	3.2%	1.8%
Biliverdin (serum)	0.030	2.2%	1.3%
Triacylglycerol (TAG 50:3(FA14:0)[*]) (serum)	0.197	1.8%	1.1%
Triacylglycerol (TAG 48:4(FA14:1)[*]) (serum)	0.121	1.7%	1.0%
Creatine (serum)	−0.078	1.6%	0.9%
Phosphatidylethanolamine (PE 18:0/20:4[*]) (serum)	−0.087	1.5%	0.9%
Phosphatidylcholine (PC 18:0/20:2[*]) (serum)	0.082	1.1%	0.6%
Xanthurenic acid (serum)	−0.201	0.8%	0.5%
Lysophosphatidylcholine (LPC 14:0[*]) (serum)	0.056	0.7%	0.4%
Phosphatidylcholine (PC 18:0/22:5[*]) (serum)	0.095	0.4%	0.3%
Triacylglycerol (TAG 52:4(FA20:2)[*]) (serum)	−0.006	0.0%	0.0%
Total		62.2%	36.1%

[#]Reordered by R^2

[*]The notation, TAG 50:5 (FA18:2) indicates that there are a total of 50 carbons with 5 double bonds in the three fatty acid chains of the lipid, of which one of the fatty acid (FA) chains has 18 carbons with 2 double bonds. For LPE 16:0, 16 indicates number of carbons and 0 indicates number of double bonds in the fatty acid chain.

TAG: triacylglycerol. LPC: lysophosphatidylcholine. CE: cholesterol ester. PE: phosphatidylethanolamine. PC: phosphatidylcholine.

Table 6.

Variables selected for predicting dietary carbohydrate intake (%E) with their R^2 and corresponding cross-validated R^2 (%) using serum + 24-hour urine metabolites without participant characteristics or established diet-related biomarkers (Model 1).

Variable[#]	Coefficient	R^2	CV-R^2
Intercept	−0.680		
Triacylglycerol (TAG 52:4(FA20:2)[*]) (serum)	0.178	21.8%	13.7%
Maltose (urine)	0.042	8.8%	5.5%
Phosphatidylcholine (PC 18:0/22:5[*]) (serum)	0.142	8.6%	5.4%
Triacylglycerol (TAG 54:1(FA20:0)[*]) (serum)	−0.103	4.8%	3.0%
3-hydroxypropionic acid (serum)	−0.104	3.7%	2.3%
Glycochenodeoxycholate (serum)	0.027	1.9%	1.2%
Lysophosphatidylcholine (LPC 22:5[*]) (serum)	0.050	2.9%	1.8%
Phosphatidylethanolamine (PE 18:0/20:4[*]) (serum)	−0.083	1.8%	1.1%
Ethylalcohol (urine)	−0.026	2.9%	1.8%
Creatine (urine)	−0.022	1.1%	0.7%
Phosphatidylcholine (PC 18:1/22:5[*]) (serum)	0.042	0.6%	0.4%
Triacylglycerol (TAG 50:4(FA14:1)[*]) (serum)	0.040	0.1%	0.1%
Cholesteryl ester (CE 20:2[*]) (serum)	0.033	0.1%	0.0%
Triacylglycerol (TAG 50:4(FA18:2)[*]) (serum)	−0.005	0.0%	0.0%
Total		59.0%	37.1%

[#]Reordered by R^2

[*]The notation, TAG 52:4 (FA20:2) indicates that there are a total of 52 carbons with 4 double bonds in the three fatty acid chains of the lipid, of which one of the fatty acid (FA) chains contains 20 carbons with 2 double bonds.

PC 18:0/22:5 indicates that the lipid has two fatty acid chains, one has 18 carbons with 0 double bond and the other has 22 carbons with 5 double bonds.

TAG: triacylglycerol. PC:phosphatidylcholine. LPC: lysophosphatidylcholine. PE: phosphatidylethanolamine. CE: cholesterol ester.