Assessment of biomarkers for carotenoids, tocopherols, retinol, vitamin B12 and folate in the Hispanic Community Health Study/Study of Latinos

Lillian A. Boe1,*, Yasmin Mossavar-Rahmani2, Daniela Sotres-Alvarez3, Martha L. Daviglus4, Ramon A. Durazo-Arvizu5, Robert C. Kaplan2,6, Pamela A. Shaw7

1 Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania
2 Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York
3 Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina
4 Institute for Minority Health Research, University of Illinois at Chicago, Chicago, Illinois
5 Biostatistics Core, The Sabin Research Institute, Keck School of Medicine, University of Southern California, Los Angeles, California
6 Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington
7 Kaiser Permanente Washington Health Research Institute, Seattle, Washington

*email: boel@pennmedicine.upenn.edu

Abstract

Measurement error is a major issue in self-reported diet that can distort diet-disease relationships. Serum biomarkers avoid the subjective bias in self-report. As part of the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), self-reported diet was collected on all participants (N=16,415). Blood concentration biomarkers for carotenoids, tocopherols, retinol, vitamin B12 and folate were collected on a subset (n=445), as part of the Study of Latinos: Nutrition and Physical Activity Assessment Study (SOLNAS). We examine the relationship between biomarker levels, self-reported intake, ethnicity and other participant characteristics in this diverse population. We build regression-based prediction equations for ten nutritional biomarkers and evaluate whether there would be sufficient precision in these prediction equations to reliably detect an association with this exposure in a multivariable Cox model. This type of predicted exposure is commonly used to assess measurement-error corrected diet-disease associations using regression calibration; however, issues of power are rarely discussed. We used simulation to study the power of detecting the association between a true average concentration marker and a hypothetical incident survival outcome in the HCHS/SOL cohort using a predicted biomarker level whose measurement characteristics were similar to those observed for SOLNAS. Good power was observed for some nutrients; whereas, a low intra-class correlation coefficient contributed to poor power for others. Repeat measures improved the ICC; however, further research is needed to understand how best to realize the potential of these dietary biomarkers. This study provides a comprehensive examination of several nutritional biomarkers within the HCHS/SOL, characterizing their associations with subject characteristics and the influence the measurement characteristics have on the power to detect associations with health outcomes.

Keywords: biomarker, design, diet, measurement error, prediction; regression calibration
Introduction

Measurement error is a major issue in self-reported diet that can distort the relationship between diet and disease outcomes\(^1\). While there are different methods for addressing measurement error, the method is dependent on how well the objective biomarker correlates with true intake and other personal characteristics, such as self-reported nutrient intake. Recovery biomarkers such as doubly labeled water (DLW) and urinary nitrogen are quantitatively related to intake and have been established as objective biomarkers of energy and protein, respectively\(^9\). In contrast to recovery biomarkers, blood concentration biomarkers such as folate, tocopherols, carotenoids and vitamin B-12 are more readily available for a wider range of nutrients\(^2\). Serum concentrations present a different challenge, as they generally cannot be related directly to absolute levels of intake. While serum concentrations do correlate with intakes of corresponding foods or nutrients, typically the correlation coefficients are much lower (less than 0.6) than that expected for recovery biomarkers\(^3\).

The measurement error in self-report of micronutrients has not been adequately investigated in at-risk populations, such as Hispanic/Latino subgroups in the United States. In the Study of Latinos: Nutrition and Physical Activity Assessment Study (SOLNAS), blood concentration biomarkers and other participant characteristics were collected from a sub-study of 476 participants from the multi-center Hispanic Community Health Study/Study of Latinos (HCHS/SOL) cohort (n = 16,415). We sought to describe how these concentration biomarkers varied with participant characteristics, such as gender, age and Hispanic ethnicity. We also sought to build prediction models for each concentration marker. These prediction models can be used to elucidate the strength of association between levels of this biomarker and the nutrient values derived from the self-reported 24-hour dietary recall. We develop these prediction equations in order to implement regression calibration, a method that replaces an unobserved true exposure (e.g. average serum concentration level) by an estimate based on other covariates (e.g. the self-reported intake and other participant characteristics). Combining data from blood concentration biomarkers with self-report data from supplement and dietary intake is a potential strategy to obtain better predictions of these objective markers of intake, which in turn could be used for estimates of diet-disease associations that are not subject to reporting bias\(^4\).

To decrease participant burden and cost, we illustrate how these models could additionally be used to impute missing concentration levels in HCHS/SOL participants, who were not part of the SOLNAS sub-study. These imputed values would allow for the serum biomarkers to be included in outcome-diet association studies in the larger cohort, similar to analyses that have been done using recovery biomarkers\(^6,7\). Such an analysis could be used to explore whether levels of this biomarker were associated with a disease outcome. We use a numerical simulation to study the statistical power of detecting an association between a true average concentration biomarker and a hypothetical time-to-event disease outcome using imputed biomarker levels based on the model characteristics observed in SOLNAS.
Subjects and methods

Study population

HCHS/SOL is a community-based cohort study of 16,415 men and women aged 18-74 years who self-identified as Hispanic/Latino and were recruited in 2008-2011 from randomly selected households in four US metropolitan areas (Bronx, NY; Chicago, IL; Miami, FL; San Diego, CA). The selected sites constitute 4 of the 11 urban centers in the US with the largest number of Hispanics/Latinos. The primary objective of HCHS/SOL is to study associations between baseline risk factors and chronic conditions in a cohort of several Hispanic/Latino backgrounds. Study participants underwent a comprehensive baseline clinical examination that included biological, behavioral, and social-demographic assessments, yearly telephone follow-up and a second clinic visit (2014-2017). Individuals were invited to participate in the SOLNAS substudy within 7 months of their HCHS/SOL visit. Participants were excluded for having any medical condition that might affect weight stability or biomarker performance, being pregnant or breastfeeding a child, weight loss or gain of > 15 pounds in past four weeks, taking medication for diabetes, or having extended travel planned during the study period. A total of 1,360 HCHS/SOL participants were invited to enroll in SOLNAS; 485 (35.7%) consented to participate. Of these 485 individuals who enrolled in SOLNAS in 2010-2012, 7 dropped out and 2 did not provide biomarker data, leaving 476 participants that completed the protocol and a subsample of 98 that repeated the study protocol after 6 months to provide reliability information. We briefly outline the study protocol here; Mossavar-Rahmani et al, 2015 provides further details.¹

Study protocol and procedures

The SOLNAS study protocol consisted of two clinic visits on all participants, with in-home activities between visits. The doubly labelled water (DLW) and urinary nitrogen recovery biomarkers were used to assess total energy expenditure and protein intake, respectively, and their data are reported elsewhere¹. At Visit 1, participants arrived following a 4-hour fast and provided a baseline urine specimen (pre-DLW spot urine sample) and completed a 24-hour dietary recall. Twelve days later, participants arrived for Visit 2 to complete the DLW protocol and collect a fasting blood draw and urine specimen. Ninety-eight participants (20%) repeated the study for visits 3 and 4. Plasma and serum samples were stored at -70 degrees until analysis.

Serum Biomarkers

Assays were completed by the central clinical chemistries laboratory at University of Minnesota Fairview Hospital Laboratory for total cholesterol, HDL, LDL, serum α-carotene, β-carotene, α-tocopherol, γ-tocopherol, vitamin B-12, β-cryptoxanthin, retinyl palmitate, folate, lycopene, retinol, and zeaxanthin. Blinded duplicates were analyzed for a 10% quality control subsample. Carotenoids and tocopherol were measured by HPLC using a slight modification of the methods described by Bieri et al, 1985, and Craft et al, 1988, which allowed for simultaneous detection of both carotenoids and tocopherol. From this method we obtained: α-tocopherol, γ-tocopherol, zeaxanthin, β-cryptoxanthin, lycopene, α-carotene, and β-carotene. Vitamin B12 was measured in serum using the Siemens ADVIA Centaur Vitamin B12 assay (Siemens Healthcare...
Diagnostics, Deerfield IL). Folate was measured using the Siemens ADVIA Centaur Folate assay (Siemens Healthcare Diagnostics, Deerfield IL). Vitamin B12 was measured using the Siemens ADVIA Centaur Vitamin B12 assay (Siemens Healthcare Diagnostics, Deerfield IL). Retinol was measured by a modification of the HPLC method described by Bieri et al, 1985, with calibration as described by Craft et al, 1988. The assay modification includes the addition of 0.015% N,N-diisopropylethylamine in the HPLC solvent as an aid to analyte recovery. Serum cholesterol measures were also analyzed by the same central laboratory from a fasting blood draw taken at HCHS/SOL study baseline. Additional details on these measures are provided in the supplementary appendix.

Dietary and participant characteristics assessment

Two 24-hour dietary recalls were collected during the HCHS/SOL baseline period; the first recall was collected in person at the clinic, and the second recall was collected by telephone within the following 3 months. In SOLNAS, an in-person 24-hour dietary recall was collected at the first visit. Further details on the administration of these procedures have been described previously. We include a few key details in the supplementary appendix. Self-reported dietary data in this analysis are based on the two-day mean of the telephone recall from the HCHS/SOL baseline period and the SOLNAS in-person 24-hour dietary recall. In the analyses, we excluded two SOLNAS participants who reported energy intake < 500 kcal/day.

At the HCHS/SOL baseline visit, participants provided data on demographic, health, lifestyle, and acculturation characteristics, including self-reported physical activity using a modified Global Physical Activity Questionnaire (GPAQ). We also recorded total 24-hour supplement intake variables corresponding to each micronutrient that may impact serum concentrations levels. In the supplementary appendix, we provide definitions for some important participant characteristic variables.

Statistical analyses

We used log-transformed serum nutrient biomarker measurements from the SOLNAS main study for α-carotene, β-carotene, α-tocopherol, γ-tocopherol, vitamin B-12, β-cryptoxanthin, retinyl palmitate, folate, lycopene, retinol, and zeaxanthin in our statistical analyses. A 2-day mean obtained from the 24-hour dietary recalls was used to estimate usual nutrient intake. The data had a missing rate of <1% for most nutrient biomarker and 24-hour dietary recall values, with the exception of vitamin B-12, folate, retinol and retinyl palmitate. Geometric means and 95% CIs (2.5th-97.5th percentiles) were reported for each nutrient variable by gender and ethnicity.

Linear regression models that predicted the underlying average biomarker level of each micronutrient were constructed by regressing the log-transformed biomarker measure on the 2-day mean of self-reported intakes and other participant characteristics. Each regression model was adjusted for the following characteristics, obtained at the baseline HCHS/SOL visit: age, gender, ethnicity, language preference, BMI, education, income, smoking status (current smoker or not), alcohol use (<1 drink/week, 1-7 drink/week, 7+ drink/week), diabetes, family history of diabetes, hypertension, high cholesterol, log total energy intake, a binary supplement use
The variables of interest for each nutrient biomarker are presented in Web Table 1. The prediction models for α-carotene, α-tocopherol, vitamin B-12, β-carotene, β-cryptoxanthin, and zeaxanthin were all adjusted for physical activity. The prediction models for α-carotene, β-carotene, β-cryptoxanthin, lycopene, and zeaxanthin were also adjusted for season. Previous literature shows that carotenoids and tocopherols need to be adjusted for cholesterol.17 Rather than using existing equations to adjust these biomarker nutrient values, we include all cholesterol-related variables in our prediction models for any carotenoids and tocopherols. Cholesterol variables consist of HDL-cholesterol (mg/dL), LDL-cholesterol (mg/dL), and lipid lowering drugs/ Antihyperlipidemics (LLD).

Web Table 2 presents a mapping between nutrient biomarkers and the self-reported nutrients that we used for each prediction model. Most biomarker nutrient measures had an exact match for a corresponding self-reported nutrient, with the exception of zeaxanthin, for which the closest corresponding self-reported measure was lutein plus zeaxanthin. Goodman et al. (1996)10 found associations between plasma α-tocopherol and self-reported dietary β-carotene, plasma β-carotene and self-reported dietary α-tocopherol, and plasma retinol and self-reported dietary β-carotene. Thus, prediction models for α-tocopherol, β-carotene, and retinol were further adjusted to include other self-reported nutrient values beyond the exact corresponding measure.

Since our statistical analysis approach was to build a prediction model for the underlying average biomarker level, we do not perform variable selection and consider the full model our gold standard model. We evaluate the predictive accuracy of our models and check for overfitting. Details on this assessment are provided in the supplementary appendix. Following our model-fitting procedure of regressing each biomarker on a nutrient-specific list of variables, we report R² values and intraclass correlation coefficients (ICC), or the correlations between repeat biomarker measures. We also compute R² values that account for the within-person variability in the biomarker, referred to as the Prentice R²;13 the correlations between the biomarker and its associated self-reported measure; and new ICC and R² measures that could exist if there were 2 and 4 repeat measures of each biomarker available on everyone in the SOLNAS substudy. The latter measures show the potential to increase the precision of the biomarker by reducing variance from day-to-day fluctuations in the diet.

We conduct simulations based on the data structure in the HCHS/SOL cohort to study the power of detecting the association between a true average biomarker concentration and a hypothetical outcome of time to incident diabetes. We assumed a Cox Proportional Hazards model and chose a hazard ratio (HR) for the true average exposure to be one that would have approximately 90% power for the HCHS/SOL simulation settings. While numerical formulas are available for power calculations in the Cox model, they are less straightforward for multivariable regression scenarios in which exposure variables may be prone to systematic and random measurement error, and thus we elected to use a simulation study instead. We simulate nutrients with high, medium, and low R² values to represent the varying predictive accuracy observed in the SOLNAS data. Details of the numerical simulation, including the assumed variable distributions, measurement error model, and outcome model, are discussed in the Supplementary Materials.
We present mean and median percent biases, average standard errors (ASE), empirical standard errors (ESE), 95% coverage probabilities (CP), and power across 1000 simulations for the estimated HR for simulated β-Cryptoxanthin, lycopene, and folate. We present results for the model using the true biomarker average level, the complete case analysis using SOLNAS data with a single observed biomarker subject to random classical measurement error (naïve biomarker), the complete case analysis using SOLNAS data only and adjusted for measurement error using regression calibration (calibrated biomarker), the main study analysis using all HCHS/SOL data with the 2-day mean 24-hour recall measure of the nutrient (naïve self-report), the main study analysis using all HCHS/SOL data with the predicted log biomarker values (calibrated self-report), and an optimal combination of the calibrated biomarker and calibrated self-report. We consider the optimal combination approach proposed by Spiegelman et al. (2001), which computes a generalized inverse-variance weighted average of the calibrated self-report and biomarker. Results for the calibrated biomarker, calibrated self-report, and optimal combination are presented for both the post-hoc (PH) regression calibration approach of Rosner et al. (1990, 1992) and the traditional regression calibration (RC) approach introduced by Prentice (1982) for time-to-event models.

Results

Baseline demographic and lifestyle characteristics of the 476 participants who completed the SOLNAS study protocol and the 16,415 HCHS/SOL parent study participants are shown in Table 1. Results indicate that SOLNAS participants are similar to the HCHS/SOL participants in age, gender, ethnicity, Spanish language preference, body mass index, education, income, family history of diabetes, hypertension, and high cholesterol. Participants were 60% female, 30% obese and had a mean age at baseline of 46.0 years in both HCHS/SOL and the SOLNAS substudy. SOLNAS participants were 29.8% Mexican, 25.8% Puerto Rican, 14.5% Cuban, 10.7% Central American, 10.1% Dominican, and 9.0% South American. The language preference of Spanish over English was 76.5% for SOLNAS and 79.9% for HCHS/SOL, with approximately 75% of both groups making under $50,000 and having less than a college education.

Table 2 presents age-adjusted geometric means for all nutritional biomarkers and 24-hour dietary recall measures by gender for the mean age in SOLNAS of 46.1 years. In Table 3, we show age- and gender-adjusted geometric means for all biomarkers and 24-hour dietary recall measures by Hispanic/Latino background for females and the mean SOLNAS age. The results suggest that there were differences in the 2-day, 24-hour dietary recall measures and the concentration biomarker measures by both gender and Hispanic/Latino background. Table 2 shows that serum biomarker levels of α-carotene and β-cryptoxanthin were quite different for females compared to males, a trend that was not present in the self-reported values. For other nutrients like retinol and zeaxanthin, only small differences were observed between females and males for both the serum biomarker level as well as self-report. Similarly, in Table 3 we observe that Hispanic/Latino backgrounds that high levels of serum concentrations sometimes also had high mean intakes according to self-report. This was true for β-carotene, where South Americans had the highest geometric mean of self-report intakes and correspondingly had large geometric means of biomarker values. For other nutrients such as lycopene, we did not see this same alignment. Mexicans, the group that self-reported the highest amount of lycopene, did not have high levels for the concentration marker of intake compared to other Hispanic/Latino backgrounds.
The fitted regression calibration (prediction) model coefficients for the logarithm of biomarker measures are shown in Table 4 for selected variables of interest. For all nutrients, Hispanic/Latino background, supplement use, age, and body mass index were important independent predictors of nutrient intake. We also observe that for a subset of nutrients including \( \alpha \)-carotene, \( \beta \)-carotene, \( \beta \)-cryptoxanthin, folate, and zeaxanthin, the corresponding self-reported measure was a useful predictor of the biomarker, as evidenced by strong associations in the multivariable model with \( p \)-values of all less than 0.03. For the remaining nutrients, we observe that the self-reported measure had very little predictive value. We check for overfitting and perform two approaches proposed by Harrel (2015)\(^1\) for variable selection when fitting reduced models. There was no conclusive evidence of variable redundancy (data not shown). As a sensitivity analysis, Web Table 3 presents the variables selected for the reduced model by stepwise selection using AIC. Reduced models for \( \alpha \)-carotene, \( \beta \)-carotene, \( \beta \)-cryptoxanthin, folate, zeaxanthin, and lycopene all kept the self-reported nutrient as a predictor after stepwise selection. This is consistent with our observations from the full models, where we saw strong associations between the self-reported nutrient and the biomarker level for the same set of nutrients, with the exception of lycopene.

In Web Table 4, we show the \( R^2 \) values based on the full regression calibration models. The \( R^2 \) values in decreasing order are as follows: \( R^2 = 0.5035 \) for \( \beta \)-cryptoxanthin, 0.4692 for \( \alpha \)-carotene, 0.4562 for \( \beta \)-carotene, 0.3486 for \( \alpha \)-tocopherol, 0.2728 for zeaxanthin, 0.2340 for \( \gamma \)-tocopherol, 0.2196 for lycopene, 0.1717 for folate, 0.1185 for retinol, and 0.1013 for vitamin B-12. We also present Prentice \( R^2 \) values and theoretical ICC and \( R^2 \) values that could be obtained if there were multiple measures of each biomarker available on all SOLNAS subjects. For nutrients such as \( \alpha \)-tocopherol with a lower ICC, we see that we can appreciably improve the \( R^2 \) by having more than one biomarker measure, as the theoretical \( R^2 \) values for two and four repeat biomarker measures are 0.4219 and 0.4700, respectively, getting closer to the ideal Prentice \( R^2 \) of 0.5304. In Web Table 4, we also provide the correlations between the log of the 2-day mean and the log-biomarker. We found that 24-hour recalls of \( \beta \)-cryptoxanthin, \( \alpha \)-carotene, \( \beta \)-carotene, and zeaxanthin were at least moderately correlated with their pertinent blood concentration markers, with correlations ranging from 0.24 to 0.41. For \( \alpha \)-tocopherol, \( \gamma \)-tocopherol, lycopene, folate, retinol, and vitamin B-12, correlations between the self-report and the biomarker ranged from 0.02 to 0.12, suggesting weak-to-little correlation due at least in part to the lower ICC. The biomarker intraclass correlation coefficients (ICC) measures, which range from 0.09 to 0.89, are discussed in further detail below.

Figures 1 and 2 show comparisons between the main and reliability study measures for the biomarker and the self-reported intakes, respectively, at visits 1 and 3. Each plot also presents the biomarker ICC measures, or the Pearson correlations between the repeat measures from the reliability subset. For the biomarkers, these correlations were \( r = 0.89 \) for \( \beta \)-cryptoxanthin, 0.87 for \( \beta \)-carotene, 0.81 for \( \alpha \)-carotene, 0.79 for Vitamin B-12, 0.70 for zeaxanthin, 0.68 for lycopene, 0.66 for \( \alpha \)-tocopherol, 0.62 for \( \gamma \)-tocopherol, 0.59 for folate, 0.17 for retinyl palmitate and 0.09 for retinol. The correlations for the self-reported 24-hour recall measurements were generally not as high: \( r = 0.14 \) for \( \beta \)-cryptoxanthin, 0.22 for \( \beta \)-carotene, 0.18 for \( \alpha \)-carotene, 0.26 for Vitamin B-12, 0.27 for lutein plus zeaxanthin, 0.07 for lycopene, 0.29 for \( \alpha \)-tocopherol, 0.27 for \( \gamma \)-tocopherol, 0.46 for folate, and 0.23 for retinol.
Table 5 presents results from our numerical study where β-cryptoxanthin, lycopene, and folate are simulated as separate predictors in a Cox proportional hazards model. The models fit using the naïve self-report that ignores measurement error as an exposure are estimated to have mean percent biases of -88.82% for β-cryptoxanthin, -98.95% for lycopene, and -91.047% for folate. We also see appreciable bias in the naïve biomarker for each nutrient, though it is never as extreme as that of the naïve self-report. The calibrated biomarker and self-report approaches show improvements in bias over the corresponding naïve approaches in all scenarios. For all nutrients, the calibrated biomarker, calibrated self-report, and optimal combination of these approaches led to absolute mean percent biases of under 5%. For β-cryptoxanthin, nominal coverage of a 95% confidence interval is maintained for all calibration approaches as well as the optimal combination. Additionally, we see a slight improvement in power from the naïve self-report (0.689) by using either the PH (0.703) or RC (0.698) optimal combination approach. For the simulations based on lycopene where the prediction model $R^2$ was 0.2196, we see similar power between the naïve self-report (0.184) and the optimal PH approach (0.180), but lower power when using the optimal RC approach (0.150). Note due to computational limitations, power for the optimal RC approach is calculated based on a Wald confidence interval computed using bootstrap standard errors, which can be problematic in finite samples due to the potential asymmetry of the bootstrap distribution\textsuperscript{22}. This likely explains the somewhat lower power observed for optimal errors. Similarly, when we simulated folate with an $R^2$ of 0.1717, neither the PH nor RC optimal approach led to increases in power over the naïve self-report. Further, simulating folate led to some CP values that were greater than the nominal 95% coverage, particularly for the RC optimal combination approach (CP=0.969).

Discussion

Diet is thought to have an important impact on the incidence of chronic diseases, but objectively assessing dietary intake is difficult. Many epidemiologic studies rely on self-reported dietary intakes as their primary exposure of interest, which is problematic due to high degree of measurement error in self-reported dietary assessment (e.g. 24-hour recalls). It is therefore of interest to consider biological indicators of nutritional intake, such as concentration biomarkers, as more objective measures of intake. Our study is the first in the United States to describe the measurement error structure of serum concentration biomarkers and corresponding self-reported 24-hour dietary recall measures in a diverse cohort of Hispanic/Latino adults. Our contribution is twofold, as we first provide a look at the levels of serum biomarkers in this cohort and we also assess how useful these biomarkers may be in the evaluation of diet-disease associations in the HCHS/SOL.

Previous studies that investigate dietary components using the HCHS/SOL cohort have been limited to nutrients such as energy and protein, for which an associated recovery biomarker measure is available. In our exploration of serum concentration biomarkers, we found levels of these markers did vary by participant characteristic. Specifically, levels of some carotenoids were higher for women compared to men, a difference that was not observed in the corresponding self-reported 2-day mean values. Similarly, we observed low serum levels of lycopene in Mexicans compared to other Hispanic/Latino backgrounds, despite the fact that this group self-reported the highest mean intake of lycopene. Serum concentration biomarkers and self-reported intakes are likely capturing different measures of diet and further, both measures
have a component of bias that makes them challenging to assess. We examine the relationship between serum biomarkers and self-reported diet and found self-reported dietary intake overall was a strong predictor of intake for α-carotene, β-carotene, β-cryptoxanthin, and zeaxanthin. For these nutrients, the self-report and corresponding biomarker were at least modestly correlated ($\rho \geq 0.24$). Overall, serum biomarkers with the exception of retinol had high ICC measures, ranging from 0.59 to 0.89, likely due to a combination of an accurately recorded assay and low biological variability. Self-reported ICC measures were much lower, ranging from 0.07 to 0.46, which is expected due to the high variability in within-person levels that can be observed in 24-hour recall measures of diet\textsuperscript{21}.

Our $R^2$ values for prediction equations built from concentration biomarkers differed slightly from previously published results using different cohorts. Lampe et al. (2016) built prediction equations for micronutrients based on a Women’s Health Initiative (WHI) feeding study and reported slightly higher $R^2$ values for α-carotene ($R^2 = 0.53$ for WHI vs. $R^2 = 0.47$ for SOLNAS), α-tocopherol (0.47 vs. 0.35), and lycopene (0.32 vs. 0.22), but substantially higher $R^2$ values for zeaxanthin (0.46 vs. 0.27), folate (0.49 vs. 0.17), and vitamin B-12 (0.51 vs. 0.10). These differences are expected, since by design feeding study participants may have lower day-to-day variability than HCHS/SOL participants. Further, some micronutrients are difficult to measure by self-report which likely contributes to the lower R-squared for SOLNAS. As an example, while true folate intake is known to be quantitatively related to its concentration biomarker\textsuperscript{9}, our prediction model relied on a highly error-prone self-reported measure of folate, which may explain the low $R^2$.

Prentice et al. (2019)\textsuperscript{14} proposed an $R^2$ cutoff value of 0.36 for identifying useful concentration biomarkers in the context of the WHI feeding study from Lampe et al. (2016). This cutoff was created based on the observation that concentration biomarkers with an $R^2 \geq 0.36$ were approximately just as correlated with their corresponding true intake values as other well-established (e.g. recovery) biomarkers were. As suggested by Prentice et al. (2019)\textsuperscript{14}, we may need to consider alternative criteria for identifying useful biomarkers in our setting. In our study, $R^2$ values ranged from 0.27 to 0.50, which are within the range of $R^2$ values observed for well-accepted recovery biomarkers of energy and protein in SOLNAS (Mossavar-Rahmani et al. 2015). We note that one important criterion for assessing concentration biomarkers is whether using the predicted usual serum level in the outcome model allows for good power to detect a diet-disease association of interest. For our exposure variable simulated to represent β-cryptoxanthin, which had a high $R^2$ of 0.5035, there was reasonable power (approximately 70%) to detect a 20% increase in serum level. The simulations of lycopene and folate indicated that as the prediction model $R^2$ values decrease, the power may fall short (< 18%). Thus, our paper offers a new benchmark for $R^2$ values that may be used to assess the usefulness of concentration biomarkers in the SOLNAS biomarker substudy context.

Our study had some limitations. Some variables used in the calibration equations, namely, HDL-cholesterol (mg/dL), LDL-cholesterol (mg/dL), Lipid lowering drugs/Antihyperlipidemtics (LLD), and total 24-hour supplement intake variables that correspond to each nutrient come from the HCHS/SOL baseline visit rather than the SOLNAS visit. Self-reported diet coincident with the serum biomarkers may lead to stronger correlations. A limitation of our simulation study is that the CP and power for the calibrated biomarker (RC) and calibrated self-report (RC) are
based on the bootstrap method, using percentile or Wald confidence intervals, which can be problematic in some finite-sample settings. Due to complexity of the two-stage regression approach in a multivariable semiparametric Cox model, there aren’t analytical formulas available to inform sample size and power calculations. Further work is needed in this area. An important finding of this study is that levels of serum concentration biomarkers often varied by participant characteristic and showed different intake patterns compared to their corresponding self-reported measures. We showed that moderate correlations between several biomarkers and their corresponding 24-hour dietary recall measures resulted in prediction models where the self-report was a useful predictor of the usual serum level. Our study also revealed that using a predicted level of the serum biomarker can be useful for estimating diet-disease associations in a time-to-event model, particularly when the $R^2$ for the prediction model is reasonably high, like for β-cryptoxanthin and α-carotene, and a cohort of similar size to HCHS/SOL. While the low $R^2$ of other micronutrients, such as α-tocopherol, led to low power for the association studies of interest, they do have potential and we hope that our paper stimulates interest in future concentration marker studies that aim to obtain repeat biomarker measures on all substudy participants. Repeat concentration markers might increase the prediction model $R^2$ and improve the power of detecting an association between a true average concentration marker and a hypothetical outcome. The authors expect that the current investigation of serum biomarkers will inform future cohort studies that explore diet-disease associations in time-to-event settings when covariate measurement error is present.

Acknowledgements

This work was supported in part by NIH grant R01-AI131771. The authors would like to thank the investigators of the HCHS/SOL study for the use of their data. A list of HCHS/SOL investigators, managers and coordinators by field center can be found here: https://sites.cscc.unc.edu/hchs/Acknowledgement.

Data Availability Statement

The data used in this paper was obtained through submission and approval of a manuscript proposal to the Hispanic Community Health Study/Study of Latinos Publications Committee, as described on the HCHS/SOL website. For more details, see https://sites.cscc.unc.edu/hchs/publications-pub.
| Characteristic                | SOLNAS N (%) | HCHS/SOL N (%) |
|------------------------------|--------------|---------------|
| Age                          |              |               |
| 18-44 years                  | 182 (38.2)   | 6701 (40.8)   |
| 45+ years                    | 294 (61.8)   | 9714 (59.2)   |
| Gender                       |              |               |
| Male                         | 187 (39.3)   | 6580 (40.1)   |
| Female                       | 289 (60.7)   | 9835 (59.9)   |
| Background                   |              |               |
| Central American             | 51 (10.7)    | 1732 (10.6)   |
| Cuban                        | 69 (14.5)    | 2348 (14.3)   |
| Dominican                    | 48 (10.1)    | 1473 (9.0)    |
| Mexican                      | 142 (29.8)   | 6472 (39.4)   |
| Puerto Rican                 | 123 (25.8)   | 2728 (16.6)   |
| South American               | 43 (9.0)     | 1072 (6.5)    |
| Other/More than one          | 0.0 (0.0)    | 503 (3.1)     |
| Language Preference          |              |               |
| English                      | 112 (23.5)   | 3296 (20.1)   |
| Spanish                      | 364 (76.5)   | 13119 (79.9)  |
| Body Mass Index (kg/m²)      |              |               |
| Underwt/normal (<25)         | 96 (20.2)    | 3321 (20.2)   |
| Overweight (25-30)           | 186 (39.1)   | 6116 (37.3)   |
| Moderately obese (30-35)     | 118 (24.8)   | 4219 (25.7)   |
| Morbidly Obese (35+)         | 76 (16.0)    | 2688 (16.4)   |
| Education                    |              |               |
| Less than High School        | 153 (32.1)   | 6207 (37.8)   |
| High School Graduate/GED ³   | 119 (24.9)   | 4180 (25.5)   |
| Some Post High School        | 70 (14.7)    | 2053 (12.5)   |
| College graduate or higher   | 135 (28.3)   | 3884 (23.7)   |
| Income                       |              |               |
| < $20,000                    | 231 (48.5)   | 7207 (43.9)   |
| $20,001-$50,000               | 137 (28.8)   | 5058 (30.8)   |
| $50,000+                     | 66 (13.9)    | 2662 (16.2)   |
| Missing                      | 42 (8.8)     | 1488 (9.1)    |
| Currently employed³          | 220 (46.2)   | 8156 (49.7)   |
| Supplement Use               | 228 (47.9)   | 7243 (44.1)   |
| Current Smoker               | 99 (20.8)    | 3166 (19.3)   |
| Alcohol                      |              |               |
| <1 drink/week                | 340 (71.4)   | 11303 (68.9)  |
| 1-7 drink/week               | 85 (17.9)    | 3255 (19.8)   |
| 7+ drink/week                | 51 (10.7)    | 1787 (10.9)   |
| Diabetes³                    | 40 (8.4)     | 3218 (19.6)   |
| Family history of Diabetes   | 211 (44.3)   | 7281 (44.4)   |
| Hypertension                 | 115 (24.2)   | 4476 (27.3)   |
| High cholesterol³            | 189 (39.7)   | 7394 (45.0)   |

1. Background=Hispanic/Latino background; GED = General Education Development; Currently employed defined as part time or full time; Diabetes defined as either fasting time > 8 and fasting glucose ≥ 126 mg/dL, fasting time ≤ 8 and fasting glucose ≥ 200 mg/dL, glucose post- oral glucose tolerance test (OGTT) ≥ 200 mg/dL, % glycated hemoglobin (A1C) ≥ 6.5%, or use of anti-diabetic medication; Hypertension defined as systolic or diastolic BP ≥ 140/90 or use of antihypertensive medications. High cholesterol defined as either total cholesterol ≥ 240 mg/dL, LDL-cholesterol ≥ 160 mg/dL, HDL-cholesterol ≥ 0 and < 40 mg/dL, or use of antihyperlipidemic medication.

Percentages may not add up to 100 because of missing data. In HCHS/SOL: ethnicity (n=87), body mass index (n=71), education (n=91), currently employed (n=306), supplement use (n=630), current smoker (n=93), alcohol (n=70), diabetes (n=21), family history of diabetes (n=117), hypertension (n=2), high cholesterol (n=21); In SOLNAS: currently employed (n=1), current smoker (n=1), family history of diabetes (n=2).
Table 2. Age-Adjusted Geometric Mean Values for Nutritional Biomarker and Self-Reported Measures of α-carotene, α-tocopherol, Vitamin B-12, β-carotene, β-cryptoxanthin, Folate, γ-tocopherol, Lycopene, Retinol, Retinyl Palmitate, and Zeaxanthin by gender, shown for the mean age in SOLNAS, 46.1 years (N=476).

| Nutrient          | Female          | Male           |
|-------------------|-----------------|----------------|
|                   | Mean  | 95% CI   | Mean  | 95% CI   |
| α-carotene        |       |          |       |          |
| 2-Day mean, 10 mcg| 5.68  | 4.49, 7.19| 5.90  | 4.40, 7.92|
| Biomarker, mcg/dL | 4.02  | 3.64, 4.44| 2.75  | 2.43, 3.12|
| α-tocopherol      |       |          |       |          |
| 2-Day mean, mg    | 6.56  | 6.12, 7.03| 8.72  | 7.99, 9.51|
| Biomarker, mg/dL  | 0.87  | 0.84, 0.90| 0.87  | 0.83, 0.90|
| B-12              |       |          |       |          |
| 2-Day mean, mcg   | 2.92  | 2.68, 3.19| 3.76  | 3.38, 4.19|
| Biomarker, 10² pg/ml | 5.60 | 5.32, 5.89| 5.12  | 4.81, 5.45|
| β-carotene        |       |          |       |          |
| 2-Day mean, 10² mcg | 8.27 | 7.15, 9.57| 8.34  | 6.96, 9.99|
| Biomarker, 10 mcg/dL | 1.29 | 1.18, 1.41| 0.82  | 0.73, 0.92|
| β-cryptoxanthin   |       |          |       |          |
| 2-Day mean, 10 mcg | 2.81  | 2.34, 3.37| 2.45  | 1.96, 3.08|
| Biomarker, mcg/dL | 8.79  | 8.01, 9.64| 6.51  | 5.80, 7.31|
| Folate            |       |          |       |          |
| 2-Day mean, 10² mcg | 3.13 | 2.95, 3.33| 4.19  | 3.89, 4.51|
| Biomarker, 10 nmol/dL | 4.31 | 4.11, 4.51| 4.18  | 3.94, 4.42|
| γ-tocopherol      |       |          |       |          |
| 2-Day mean, 10 mg | 0.71  | 0.65, 0.77| 0.96  | 0.86, 1.07|
| Biomarker, mg/dL  | 0.19  | 0.18, 0.19| 0.17  | 0.16, 0.18|
| Lycopene          |       |          |       |          |
| 2-Day mean, 10² mcg | 2.90 | 2.06, 4.08| 5.43  | 3.54, 8.31|
| Biomarker, 10 mcg/dL | 3.08 | 2.91, 3.25| 3.26  | 3.04, 3.49|
| Retinol           |       |          |       |          |
| 2-Day mean, 10² mcg | 2.15 | 1.92, 2.41| 2.71  | 2.35, 3.11|
| Biomarker, mcg/dL | 0.49  | 0.47, 0.51| 0.56  | 0.53, 0.59|
| Retinyl palmitate |       |          |       |          |
| 2-Day mean, 10² mcg | 0.11 | 0.08, 0.15| 0.10  | 0.08, 0.14|
| Biomarker, mcg/dL |       |          |       |          |
| Zeaxanthin        |       |          |       |          |
| 2-Day mean, 10² mcg | 4.94 | 4.35, 5.61| 5.34  | 4.56, 6.26|
| Biomarker, 10 mcg/dL | 1.71 | 1.61, 1.81| 1.78  | 1.65, 1.91|

1. N=476 for biomarker means and N=474 for self-reported measures due to two 24-hour recalls that listed an energy intake of <500 kcal/day and were deleted. 2. No exact 24-hour recall corresponding measure existed in data.

Missing data for each assay are as follows: B-12 (n=5), Retinyl palmitate (n=221), Total Folate (n=11), Retinol (n=18).
| Nutrient                      | Central American       | Cuban          | Dominican       | Mexican         | Puerto Rican    | South American  |
|------------------------------|------------------------|----------------|----------------|----------------|----------------|----------------|
|                             | Mean 95% CI            | Mean 95% CI    | Mean 95% CI    | Mean 95% CI    | Mean 95% CI    | Mean 95% CI    |
| **α-carotene**               |                        |                |                |                |                |                |
| 2-Day mean, 10² mcg          | 0.69 0.39–1.23         | 0.69 0.41–1.15 | 0.91 0.50–1.63 | 0.59 0.42–0.84 | 0.28 0.19–0.42 | 1.13 0.61–2.09 |
| Biomarker, mcg/dL            | 5.51 4.42–6.87         | 4.18 3.43–5.09 | 8.27 6.60–10.36 | 3.96 3.46–4.54 | 2.26 1.95–2.63 | 5.83 4.60–7.39 |
| **α-tocopherol**             |                        |                |                |                |                |                |
| 2-Day mean, mg               | 6.09 5.15–7.20         | 6.48 5.57–7.52 | 4.45 3.74–5.29 | 7.86 7.09–8.72 | 6.18 5.52–6.92 | 6.86 5.73–8.22 |
| Biomarker, mg/dL             | 0.85 0.79–0.92         | 0.83 0.77–0.89 | 0.87 0.80–0.94 | 0.94 0.90–0.99 | 0.81 0.76–0.85 | 0.85 0.78–0.93 |
| **B-12**                     |                        |                |                |                |                |                |
| 2-Day mean mcg               | 2.57 2.08–3.18         | 2.66 2.21–3.22 | 2.34 1.88–2.91 | 3.39 2.98–3.86 | 2.97 2.57–3.42 | 2.77 2.20–3.48 |
| Biomarker, 10⁵ pg/ml         | 5.59 4.95–6.31         | 4.83 4.33–5.38 | 5.09 4.49–5.76 | 6.22 5.76–6.71 | 5.44 5.01–5.90 | 5.79 5.08–6.60 |
| **β-carotene**               |                        |                |                |                |                |                |
| 2-Day mean, 10³ mcg          | 0.87 0.61–1.24         | 0.79 0.58–1.09 | 0.85 0.59–1.22 | 1.04 0.83–1.29 | 0.55 0.43–0.70 | 1.08 0.74–1.59 |
| Biomarker, 10 mcg/dL         | 1.34 1.09–1.66         | 1.17 0.97–1.41 | 1.78 1.43–2.21 | 1.60 1.41–1.82 | 0.81 0.70–0.93 | 1.60 1.27–2.01 |
| **β-cryptoxanthin**          |                        |                |                |                |                |                |
| 2-Day mean, 10 mcg           | 3.09 2.03–4.72         | 1.52 1.04–2.22 | 2.12 1.37–3.29 | 5.51 4.24–7.15 | 1.51 1.13–2.01 | 3.83 2.43–6.04 |
| Biomarker, 10 mcg/dL         | 0.85 0.70–1.03         | 0.64 0.54–0.76 | 0.93 0.76–1.13 | 1.48 1.31–1.66 | 0.47 0.41–0.53 | 1.16 0.94–1.42 |
| **Folate**                   |                        |                |                |                |                |                |
| 2-Day mean, 10² mcg          | 3.10 2.68–3.58         | 3.08 2.70–3.50 | 2.61 2.16–2.91 | 2.56 2.35–2.89 | 2.86 2.59–3.16 | 3.40 2.90–3.97 |
| Biomarker, 10 nmol/dL        | 4.27 3.82–4.78         | 3.59 3.24–3.98 | 4.26 3.80–4.78 | 4.60 4.29–4.93 | 4.31 3.99–4.65 | 4.43 3.92–4.99 |
| **γ-tocopherol**             |                        |                |                |                |                |                |
| 2-Day mean mg                | 6.92 5.58–8.59         | 7.12 5.87–8.63 | 4.49 3.59–5.61 | 8.55 7.49–9.76 | 6.86 5.93–7.93 | 6.57 5.21–8.28 |
| Biomarker mg/dL              | 0.18 0.16–0.20         | 0.19 0.18–0.21 | 0.16 0.15–0.18 | 0.21 0.19–0.22 | 0.17 0.16–0.18 | 0.19 0.17–0.21 |
| **Lycopene**                 |                        |                |                |                |                |                |
| 2-Day mean, 10² mcg          | 2.08 0.89–4.85         | 3.06 1.43–6.51 | 1.42 0.59–3.40 | 4.09 2.43–6.89 | 2.71 1.53–4.79 | 3.04 1.22–7.54 |
| Biomarker, 10 mcg/dL         | 2.79 2.44–3.20         | 3.25 2.88–3.67 | 4.07 3.54–4.67 | 2.95 2.71–3.20 | 3.08 2.81–3.38 | 2.69 2.33–3.12 |
| **Retinol**                  |                        |                |                |                |                |                |
| 2-Day mean, 10² mcg          | 1.79 1.29–2.24         | 1.90 1.48–2.43 | 1.73 1.30–2.30 | 2.53 2.14–3.00 | 2.28 1.90–2.75 | 2.01 1.49–2.70 |
| Biomarker, mg/dL             | 0.49 0.44–0.54         | 0.54 0.49–0.59 | 0.47 0.42–0.52 | 0.50 0.46–0.53 | 0.46 0.43–0.49 | 0.48 0.43–0.54 |
| **Retinyl palmitate**        | 0.12 0.06–0.22         | 0.17 0.10–0.28 | 0.03 0.01–0.06 | 0.22 0.15–0.31 | 0.04 0.03–0.06 | 0.10 0.06–0.19 |
| 2-Day mean mcg               | 5.45 4.05–7.34         | 3.43 2.63–4.48 | 2.99 2.20–4.07 | 7.70 6.41–9.25 | 3.55 2.91–4.35 | 6.34 4.60–8.74 |
| Biomarker, mg/dL             | 1.86 1.63–2.14         | 1.52 1.34–1.71 | 1.93 1.68–2.22 | 2.00 1.84–2.18 | 1.31 1.19–1.43 | 1.84 1.59–2.13 |

1. N=476 for biomarker means and N=474 for self-reported measures due to two 24-hour recalls that listed an energy intake of <500 kcal/day and were deleted. 2. No exact 24-hour recall corresponding measure existed in data.

Missing data for each assay are as follows: B-12 (n=5), Retinyl palmitate (n=221), Total Folate (n=11), Retinol (n=18).
Table 4. Regression calibration β coefficients for logarithm of biomarker using self-reported intake and other important subject characteristics. Regressions are done on the log-transformed measures (N=4761).

| Variable         | α-carotene | α-tocopherol | VB12 | β-carotene | β-cryptoxanthin |
|------------------|------------|--------------|------|------------|-----------------|
| Intercept        | 0.051      | 0.279        | -0.668 | 0.109    | 6.428           | 1.085         | 1.287 | 0.247 |
| Male             |            |              |       |            |                 |               |       |       |
| Female           | 0.096      | 0.017        | 0.02  | 0.022      | 0.032           | 0.031         | 0.105 | 0.030 |
| Male: Female     | -0.045     | 0.017        | 0.002 | 0.018      | 0.018           | 0.018         | 0.007 | 0.019 |
| Male: Female: Female | -0.045 | 0.017 | 0.002 | 0.018 | 0.018 | 0.018 | 0.007 | 0.019 |

| Variable         | β          | SE           | β          | SE           | β          | SE           | β          | SE           |
|------------------|------------|--------------|------------|--------------|------------|--------------|------------|--------------|
| Intercept        | 3.049      | 0.240        | -2.115     | 0.120        | 2.472      | 0.200        | -0.906     | 0.180        | 2.087      | 0.229        |
| Male             |            |              |            |              |            |              |            |              |            |              |
| Female           | 0.090      | 0.039        | 0.015      | 0.022        | 0.011      | 0.008        | -0.001     | 0.019        | 0.084      | 0.022        |
| Male: Female     | -0.076     | 0.07         | -0.155     | 0.057        | -0.078     | 0.091        | -0.021     | 0.064        | -0.083     | 0.089        |
| Male: Female: Female | -0.076 | 0.07 | -0.155 | 0.057 | -0.078 | 0.091 | -0.021 | 0.064 | -0.083 | 0.089 |

| Variable         | β          | SE           | β          | SE           | β          | SE           | β          | SE           |
|------------------|------------|--------------|------------|--------------|------------|--------------|------------|--------------|
| Intercept        | 0.334      | 0.124        | -0.095     | 0.043        | -0.115     | 0.082        | -0.18      | 0.129        | -0.501     | 0.108        |
| Central Am.      |            |              |            |              |            |              |            |              |            |              |
| Cuban            | 0.174      | 0.115        | -0.109     | 0.042        | -0.226     | 0.077        | -0.124     | 0.117        | -0.574     | 0.104        |
| Dominican        | 0.780      | 0.137        | -0.066     | 0.046        | -0.213     | 0.085        | 0.146      | 0.14         | -0.339     | 0.12         |
| Puerto Rican     | -0.208     | 0.099        | -0.099     | 0.035        | -0.117     | 0.066        | -0.391     | 0.102        | -0.783     | 0.089        |
| South Am.        | 0.322      | 0.13         | -0.124     | 0.046        | -0.071     | 0.084        | -0.06      | 0.132        | -0.247     | 0.113        |
| BMI6             | -0.018     | 0.006        | 0.002      | 0.002        | -0.003     | 0.004        | -0.022     | 0.006        | -0.013     | 0.005        |
| Age7             | 0.009      | 0.003        | 0.006      | 0.001        | -0.001     | 0.002        | 0.007      | 0.003        | 0.000      | 0.003        |
| Female           | 0.332      | 0.085        | -0.012     | 0.03         | 0.04       | 0.053        | 0.41       | 0.086        | 0.142      | 0.074        |
| Supp. Use9       | 0.087      | 0.069        | 0.045      | 0.025        | 0.104      | 0.046        | 0.149      | 0.071        | 0.085      | 0.060        |

| Variable         | β          | SE           | β          | SE           | β          | SE           | β          | SE           |
|------------------|------------|--------------|------------|--------------|------------|--------------|------------|--------------|
| Intercept        | 906        | 0.180        | 5.180      | 2.280        | 906        | 0.180        | 5.180      | 2.280        |
| Self-report      |            |              |            |              |            |              |            |              |
| Background       | 0.009      | 0.003        | 0.006      | 0.001        | 0.007      | 0.003        | 0.000      | 0.003        |
| Central Am.      | -0.076     | 0.07         | -0.155     | 0.057        | -0.078     | 0.091        | -0.021     | 0.064        | -0.083     | 0.089        |
| Cuban            | -0.182     | 0.067        | -0.105     | 0.055        | 0.061      | 0.082        | 0.093      | 0.061        | -0.157     | 0.082        |
| Dominican        | 0.013      | 0.071        | -0.214     | 0.061        | 0.39       | 0.096        | -0.089     | 0.069        | -0.039     | 0.096        |
| Puerto Rican     | -0.013     | 0.057        | -0.177     | 0.047        | 0.099      | 0.071        | -0.061     | 0.053        | -0.24      | 0.071        |
| South Am.        | -0.019     | 0.072        | -0.116     | 0.061        | -0.144     | 0.093        | -0.074     | 0.067        | -0.123     | 0.091        |
| BMI6             | -0.01      | 0.003        | 0.008      | 0.003        | -0.004     | 0.004        | -0.002     | 0.003        | -0.009     | 0.004        |
| Age7             | 0.004      | 0.002        | 0.001      | 0.002        | -0.003     | 0.002        | 0.001      | 0.002        | 0.001      | 0.002        |
| Female           | 0.058      | 0.045        | 0.055      | 0.040        | 0.017      | 0.059        | -0.065     | 0.045        | 0.004      | 0.059        |
| Supp. Use9       | 0.109      | 0.042        | -0.096     | 0.033        | -0.034     | 0.050        | 0.072      | 0.037        | -0.017     | 0.048        |

1. Calibration models for each nutrient were fit on N=452 subjects or fewer due to the following missing variables in the SOLNAS data: current smoker (N=1), family history of diabetes (N=2), vitamin B-12 biomarker (N=5), retinyl palmitate biomarker (N=221), folate biomarker (N=11), retinol biomarker (N=18), LLD medication (N=10), HDL cholesterol (N=1), LDL cholesterol (N=12), vitamin B-12 supplement intake (N=53), β-carotene supplement intake (N=49), folate supplement intake (N=49), lycopene supplement intake (N=49), zeaxanthin supplement intake (N=49). 2. The intercept represents the mean for the baseline group of nonsmoking, diabetes and hypertension-free males of average age (46.1 years), average body mass index of 29.6, Spanish language preference, Mexican background, low income (<$20,000), low alcohol use (<1 drink/week), and less than high school education who do not have high cholesterol, have no family history of diabetes, and who do not take supplements.3. Hispanic/Latino Background 4. Central American 5. South American 6 Body Mass Index (kg/m²), centered at mean (29.6) 7. Age is centered at mean 46.1 years, 8. Supplement Use
Table 5. Mean and median % bias, average standard errors (ASE), empirical standard errors (ESE), coverage probabilities (CP), and power for β-cryptoxanthin, Lycopene, and Folate in a time-to-event model when the true β corresponding to the nutrient measure in the outcome model is assumed to be \( \log(0.862) = -0.149 \) for β-cryptoxanthin, \( \log(0.451) = -0.796 \) for Lycopene, and \( \log(0.651) = -0.429 \) for Folate, which correspond to hazard ratios for a 20% increase in true nutrient consumption of 0.973, 0.865, and 0.925, respectively. Results are based on 2500 simulations and 1000 bootstrap replications.

| β-cryptoxanthin | Mean % Bias | Median % Bias | ASE\(^1\) | ESE\(^2\) | CP | Power |
|-----------------|------------|--------------|----------|----------|----|-------|
| Truth           | 0.279      | 0.481        | 0.076    | 0.075    | 0.952 | 0.921 |
| Naïve Biomarker | -12.755    | -9.238       | 0.426    | 0.432    | 0.951 | 0.079 |
| Calibrated Biomarker (PH\(^3\)) | 1.893 | 5.612 | 0.497 | 0.506 | 0.953 | 0.079 |
| Calibrated Biomarker (RC\(^4\)) | 1.818 | 5.886 | 0.524 | 0.507 | 0.948 | 0.082 |
| Naïve Self-Report | -88.82 | -88.933 | 0.012 | 0.012 | 0.000 | 0.689 |
| Calibrated Self-Report (PH\(^3\)) | 0.151 | -0.331 | 0.106 | 0.105 | 0.950 | 0.684 |
| Calibrated Self-Report (RC\(^4\)) | 0.151 | -0.331 | 0.107 | 0.105 | 0.948 | 0.679 |
| Optimal (PH\(^3\)) | 0.134 | -0.555 | 0.104 | 0.105 | 0.943 | 0.703 |
| Optimal (RC\(^4\)) | 0.061 | -0.671 | 0.104 | 0.105 | 0.946 | 0.698 |

| Lycopene | Mean % Bias | Median % Bias | ASE | ESE | CP | Power |
|----------|------------|--------------|-----|-----|----|-------|
| Truth    | 0.515      | 0.948        | 0.198 | 0.197 | 0.950 | 0.899 |
| Naïve Biomarker | -33.613 | -29.26 | 0.988 | 1.013 | 0.945 | 0.074 |
| Calibrated Biomarker (PH\(^3\)) | -1.722 | 2.352 | 1.467 | 1.509 | 0.953 | 0.073 |
| Calibrated Biomarker (RC\(^4\)) | -1.448 | 2.623 | 1.583 | 1.522 | 0.947 | 0.074 |
| Naïve Self-Report | -98.952 | -98.973 | 0.006 | 0.006 | 0.000 | 0.184 |
| Calibrated Self-Report (PH\(^3\)) | 3.357 | -1.302 | 0.649 | 0.657 | 0.962 | 0.153 |
| Calibrated Self-Report (RC\(^4\)) | 3.357 | -1.302 | 0.997 | 0.657 | 0.952 | 0.178 |
| Optimal (PH\(^3\)) | -0.085 | -2.580 | 0.585 | 0.584 | 0.952 | 0.18 |
| Optimal (RC\(^4\)) | -2.100 | -3.186 | 0.624 | 0.589 | 0.963 | 0.15 |

| Folate | Mean % Bias | Median % Bias | ASE | ESE | CP | Power |
|--------|------------|--------------|-----|-----|----|-------|
| Truth  | 0.254      | 0.425        | 0.126 | 0.125 | 0.950 | 0.904 |
| Naïve Biomarker | -47.812 | -45.279 | 0.56 | 0.581 | 0.929 | 0.069 |
| Calibrated Biomarker (PH\(^3\)) | -1.813 | -1.076 | 1.059 | 1.108 | 0.945 | 0.068 |
| Calibrated Biomarker (RC\(^4\)) | -1.288 | -2.872 | 1.253 | 1.119 | 0.944 | 0.072 |
| Naïve Self-Report | -91.047 | -91.175 | 0.037 | 0.037 | 0.000 | 0.167 |
| Calibrated Self-Report (PH\(^3\)) | 4.986 | -2.230 | 0.457 | 0.476 | 0.968 | 0.116 |
| Calibrated Self-Report (RC\(^4\)) | 4.986 | -2.230 | 2.116 | 0.476 | 0.951 | 0.166 |
| Optimal (PH\(^3\)) | -0.867 | -3.788 | 0.407 | 0.402 | 0.959 | 0.142 |
| Optimal (RC\(^4\)) | -3.132 | -7.529 | 0.465 | 0.415 | 0.969 | 0.099 |

1. ASE is defined as the mean of the estimated standard errors from the model or bootstrap standard errors. For the RC approaches, ASEs are calculated as bootstrap standard errors while coverage probability and power are calculated from a bootstrap percentile confidence interval. 2. ESE is the empirical standard deviation of the estimated coefficients across simulations. 3. PH = Post-Hoc Regression Calibration type approach. 4. RC = Traditional Regression Calibration type approach.
Figure 1. Comparison of the main study and reliability measures for each biomarker measurement (N=95). The dotted line denotes the 45-line (y=x). Each plot gives Pearson correlations for the logarithm of Visit 1 measures versus Visit 3 measures.
Figure 2. Comparison of the main study and reliability measures for each self-reported 24-hour Recall measurement (N=95). The dotted line denotes the 45-line (y=x). Each plot gives Pearson correlations for the logarithm of Visit 1 measures versus Visit 3 measures.
References

1. Mossavar-Rahmani Y, Shaw PA, Wong WW, Sotres-Alvarez D, Gellman MD, Van Horn L, Stoutenberg M, Daviglus ML, Wylie-Rosett J, Siega-Riz AM, Ou FS. Applying recovery biomarkers to calibrate self-report measures of energy and protein in the Hispanic Community Health Study/Study of Latinos. American journal of epidemiology. 2015 May 20;181(12):996-1007.

2. Mossavar-Rahmani Y, Sotres-Alvarez D, Wong WW, Loria CM, Gellman MD, Van Horn L, Alderman MH, Beasley JM, Lora CM, Siega-Riz AM, Kaplan RC, Shaw PA. Applying recovery biomarkers to calibrate self-report measures of sodium and potassium in the Hispanic Community Health Study/Study of Latinos. Journal of human hypertension. 2017 Jul;31(7):462.

3. Mann Jim, and Stewart Truswell, eds. Essentials of human nutrition. Oxford University Press, 2017.

4. Pfeiffer CM, Hughes JP, Lacher DA, Bailey RL, Berry RJ, Zhang M, et al. Estimation of trends in serum and rbc folate in the U.S. Population from pre- to postfortification using assay-adjusted data from the nhanes 1988-2010. J Nutr. 2012;142:886-893.

5. Freedman LS, Kipnis V, Schatzkin A, Tasevska N, Potischman N. Can we use biomarkers in combination with self-reports to strengthen the analysis of nutritional epidemiologic studies? Epidemiol Perspect Innov. 2010;7:2.

6. Prentice RL, Shaw PA, Bingham S, and others. Biomarker-calibrated energy and protein consumption and increased cancer risk among postmenopausal women. Am J Epidemiol. 2009;169(8):977-989.

7. Prentice RL, Huang Y, Kuller LH, and others. Biomarker-calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women, Epidemiol 2011; 22:170-179.

8. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, Wood AM, Carpenter JR. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. Bmj. 2009 Jun 29;338:b2393.

9. Lampe JW, Huang Y, Neuhausser ML, Tinker LF, Song X, Schoeller DA, Kim S, Raftery D, Di C, Zheng C, Schwarz Y. Dietary biomarker evaluation in a controlled feeding study in women from the Women’s Health Initiative cohort. The American journal of clinical nutrition. 2016 Dec 28;105(2):466-75.

10. Goodman GE, Thornquist M, Kestin M, Metch B, Anderson G, Omenn GS. The association between participant characteristics and serum concentrations of beta-carotene, retinol, retinyl palmitate, and alpha-tocopherol among participants in the Carotene and Retinol Efficacy Trial (CARET) for prevention of lung cancer. Cancer Epidemiology and Prevention Biomarkers. 1996 Oct 1;5(10):815-21.

11. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015 Aug 14.

12. Nierenberg DW, Stukel TA, Baron JA, Dain BJ, Greenberg ER, Skin Cancer Prevention Study Group. Determinants of plasma levels of beta-carotene and retinol. American Journal of Epidemiology. 1989 Sep 1;130(3):511-21.

13. Prentice RL, Mossavar-Rahmani Y, Huang Y, Van Horn L, Beresford SA, Caan B, Tinker L, Schoeller D, Bingham S, Eaton CB, Thomson C. Evaluation and comparison of
food records, recalls, and frequencies for energy and protein assessment by using recovery biomarkers. American journal of epidemiology. 2011 Jul 15;174(5):591-603.

14. Prentice RL, Pettinger M, Neuhouser ML, Tinker LF, Huang Y, Zheng C, Manson JE, Mossavar-Rahmani Y, Anderson GL, Lampe JW. Application of blood concentration biomarkers in nutritional epidemiology: example of carotenoid and tocopherol intake in relation to chronic disease risk. The American journal of clinical nutrition. 2019 Mar 27.

15. Shaw PA, McMurray R, Butte N, Sotres-Alvarez D, Sun H, Stoutenberg M, Evenson KR, Wong WW, Moncrief AE, Sanchez-Johnsen LA, Carnethon MR. Calibration of activity-related energy expenditure in the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). Journal of science and medicine in sport. 2019 Mar 1;22(3):300-6.

16. Wang L, Gaziano JM, Norkus EP, Buring JE, Sesso HD. Associations of plasma carotenoids with risk factors and biomarkers related to cardiovascular disease in middle-aged and older women. The American journal of clinical nutrition. 2008 Sep 1;88(3):747-54.

17. Gross M, Yu X, Hannan P, Prouty C, Jacobs Jr DR. Lipid standardization of serum fat-soluble antioxidant concentrations: the YALTA study. The American journal of clinical nutrition. 2003 Feb 1;77(2):458-66.

18. Spiegelman D, Carroll RJ, Kipnis V. Efficient Regression Calibration for Logistic Regression in Main Study/Internal Validation Study Designs with an Imperfect Reference Instrument.

19. Prentice RL. Covariate measurement errors and parameter estimation in a failure time regression model. Biometrika. 1982 Aug 1;69(2):331-42.

20. Frank E Harrell Jr (2019). rms: Regression Modeling Strategies. R package version 5.1-4. https://CRAN.R-project.org/package=rms

21. Willett W. Nutritional epidemiology. Oxford university press; 2012 Nov 7.

22. Efron B. Better bootstrap confidence intervals. Journal of the American statistical Association. 1987 Mar 1;82(397):171-85.
Web Appendix

Supplemental Methods for "Prediction equations for serum/plasma for carotenoids, tocopherols, retinol, vitamin B12 and folate in parent study using SOLNAS data"

Authors: Boe L, Mossavar-Rahmani Y, Sotres-Alvarez D, Daviglus M, Durazo-Arvizu R, Kaplan R, Shaw PA

Supplemental Details: Study protocol and procedures

Serum Biomarkers

As described in the main paper, serum cholesterol measures were analyzed by the same central laboratory from a fasting blood draw taken at HCHS/SOL Total cholesterol was measured by a Roche Modular P Chemistry Analyzer (Roche Diagnostics Corporation) using a cholesterol oxidase enzymatic method (Roche Diagnostics, Indianapolis, IN 46250). HDL-Cholesterol was measured in serum on a Roche Modular P Chemistry Analyzer (Roche Diagnostics Corporation) using a direct magnesium/dextran sulfate method (Roche Diagnostics, Indianapolis, IN 46250) and LDL was derived analytically.

Dietary and participant characteristics assessment

24-hour dietary recalls in this study were conducted using the Nutrition Data System for Research software (version 11) developed by the Nutrition Coordinating Center, University of Minnesota, (Minneapolis, Minnesota) and were led by bilingual interviewers using the language that each respondent preferred. The telephone recalls from the HCHS/SOL baseline period and the SOLNAS in-person 24-hour dietary recall were used to create the self-reported two-day mean in this analysis because these data were recorded at the closest point in time to the serum biomarkers (see Web Figure 1).

We now define a few key participant characteristics in the HCHS/SOL cohort. Participants were defined as supplement users (yes/no) if at least one supplement was reported in the past 30 days from the 30-day NDSR Dietary Supplement Assessment Module conducted at HCHS/SOL baseline clinic visit. Hypertension was determined to be present if a systolic/diastolic BP of greater than or equal to 140/90 was recorded or if the participant reported use of antihypertensive medications. Participants were recorded as having high cholesterol if any of the following were true: total cholesterol ≥ 240 mg/dL, LDL-cholesterol ≥ 160 mg/dL, HDL-cholesterol ≥ 0 and < 40 mg/dL, or participant reported use of antihyperlipidemic medication.

Procedures to Check for Overfitting

We consider two methods proposed by Harrell (2015)\(^1\) for variable selection to fit reduced models in the presence of overfitting. The first is a redundancy analysis procedure for constructing reduced calibration models for each nutrient based on how well a variable can be predicted from other variables in the model. The procedures are implemented using the \texttt{redund} function in the \texttt{Hmisc} package in \texttt{R}.\(^2\) This process expands continuous predictors into restricted cubic spline basis functions and expands categorical predictors into dummy variables. Ordinary
least squares is then used to predict each individual predictor with all components of the remaining predictors. Any variable that can be predicted with a high $R^2$ value by the remaining set of predictors is removed. We continue this process for all other predictors until no single predictor can be predicted with an $R^2$ of at least $r$ or until dropping a predictor causes a variable that was dropped earlier to no longer be able to be predicted at level $r$ from the reduced list of variables. We choose $r = 0.80$ for our threshold.

A second procedure for data reduction is variable clustering, which can be used to determine the relationship between candidate predictor variables. Variable clustering is an approach to assessing collinearity and redundancy as well as creating clusters of variables that may be treated as a single variable. The `Hmisc varclus` function in R performs a hierarchical cluster analysis on variables using a similarity matrix. For a similarity measure, we select the robust Hoeffding’s D statistic which has the ability to identify dependencies between any two variables. Hoeffding’s D provides a measure of the agreement between $F(x,y)$ and $G(x)H(y)$, where $F$ is the joint cumulative distribution function (CDF) between two variables $X$ and $Y$ and $G$ and $H$ are marginal CDFs.

**Details of Numerical Simulation**

Simulations were based on the HCHS/SOL cohort. We chose a population of similar size, with $N = 16,415$ for the full cohort, $N = 476$ for the SOLNAS substudy, and $N = 95$ in the reliability study for which we have two biomarker measures available. We consider three separate outcome models for three nutrients: $\beta$-Cryptoxanthin, lycopene, and folate. For each model, we further assume that age and body mass index (BMI) are related to the outcome of interest. We generate three multivariate normal distributions, one for each nutrient, with means and covariance matrices based on the full HCHS/SOL cohort. For example, we used a mean of the log of the 2-day average of 24-hour recall measures of $\beta$-Cryptoxanthin, lycopene, and folate of 3.261392, 5.605585, and 5.736064, respectively, and a mean age of 45.81989 and mean BMI of 29.77589 to generate this data. Below are the covariance matrices used to generate our data, with the variance for the corresponding nutrient in position $[1,1]$, the variance for age in position $[2,2]$, and the variance for BMI in position $[3,3]$:  

$$
\Sigma_{\beta\text{-crypt}} = \begin{bmatrix}
2.7095730 & 0.5280317 & 0.4143209 \\
0.5280317 & 194.0924000 & 8.3544090 \\
-0.4143209 & 8.3544090 & 36.8888900
\end{bmatrix},
$$

$$
\Sigma_{\text{Lycopene}} = \begin{bmatrix}
9.0749150 & -2.981719 & -0.6255943 \\
-2.9817190 & 194.0924000 & 8.3544090 \\
-0.6255943 & 8.3544090 & 36.8888900
\end{bmatrix},
$$

and

$$
\Sigma_{\text{Folate}} = \begin{bmatrix}
0.2948574 & -0.6786461 & -0.2807268 \\
-0.6786461 & 194.0924000 & 8.3544090 \\
-0.2807268 & 8.3544090 & 36.8888900
\end{bmatrix}.
$$
To simulate the true nutrient measure from the simulated error-prone measure, we assumed the following model: \( X = \alpha_0 + \alpha_1 X^* + \alpha_2 \text{AGE} + \alpha_3 \text{BMI} + u \), where \( X \) is the true nutrient measure, \( X^* \) is the error-prone measure generated from the multivariate normal distribution described above, and \( u \) is mean zero, random error with variance \( \sigma_u^2 \). The \( \alpha_0, \alpha_1, \alpha_2, \) and \( \alpha_3 \) values used to simulate each true nutrient measure were based on the calibration coefficients presented in Table 4 of the main manuscript. To simulate the biomarker measure of each nutrient, we assumed the classical measurement error model as follows: \( X'' = X + \epsilon \), where \( X'' \) is the biomarker measure, \( X \) is the true nutrient measure simulated above, and \( \epsilon \) is mean zero, random error with variance \( \sigma_\epsilon^2 \). We simulate the R\(^2 \) values based on the full prediction models in the data to be 0.5034792 for \( \beta \text{Cryptoxanthin} \), 0.2196337 for Lycopene, and 0.1716752 for Folate. This is accomplished by simulating a total variance \( \sigma_T^2 = \sigma_\epsilon^2 + \sigma_u^2 \) as a function of the R\(^2 \). Note that the R\(^2 \) for each prediction model has the following form:

\[
R^2 = \frac{\text{Var}(X^\ast) - \sigma_T^2}{\text{Var}(X^\ast)}
\]

We can rearrange this formula to solve for \( \sigma_T^2 \) as follows: \( \sigma_T^2 = (1-R^2)\text{Var}(X^\ast) \). Lastly, to simulate \( \sigma_T^2 \), we must derive the variance of the biomarker, \( X^\ast \), as below:

\[
\text{Var}(X^\ast) = \text{Var}(\alpha_0 + \alpha_1 X^* + \alpha_2 \text{AGE} + \alpha_3 \text{BMI} + u + \epsilon) \\
= \alpha_1^2 \text{Var}(X^*) + \alpha_2^2 \text{Var}(\text{AGE}) + \alpha_3^2 \text{Var}(\text{BMI}) + \sigma_\epsilon^2 + 2\alpha_1\alpha_2 \text{Cov}(X^*, \text{AGE}) \\
+ 2\alpha_1\alpha_3 \text{Cov}(X^*, \text{BMI}) + 2\alpha_2\alpha_3 \text{Cov}(\text{AGE}, \text{BMI}) = A + \sigma_T^2
\]

where \( A = \alpha_1^2 \text{Var}(X^*) + \alpha_2^2 \text{Var}(\text{AGE}) + \alpha_3^2 \text{Var}(\text{BMI}) + 2\alpha_1\alpha_2 \text{Cov}(X^*, \text{AGE}) \\
+ 2\alpha_1\alpha_3 \text{Cov}(X^*, \text{BMI}) + 2\alpha_2\alpha_3 \text{Cov}(\text{AGE}, \text{BMI}) \)

Now, we see that \( \sigma_T^2 = (1-R^2)\text{Var}(X^\ast) = (1-R^2)(A + \sigma_T^2) \). Rearranging, we have the following:

\[
\sigma_T^2 = \frac{A(1-R^2)}{R^2}
\]

which is used to simulate the R\(^2 \) values for the three regression calibration equations based on those observed in the data. We choose \( \sigma_\epsilon^2 \) based on the variance of repeat biomarker measures observed in the reliability subset. In particular, we found the ratio between the variance of repeat biomarker measures (\( \sigma_{\epsilon,\text{Data}}^2 \)) and the prediction model residual variance from the data (\( \sigma_{T,\text{Data}}^2 \)), and solved the equation below for \( \sigma_\epsilon^2 \), where the value \( \sigma_{T,\text{SIM1}}^2 \) was chosen based on the simulated value of \( \sigma_T^2 \) described above, but calculated from a single simulation iteration:

\[
\frac{\sigma_{\epsilon,\text{Data}}^2}{\sigma_{T,\text{Data}}^2} = \frac{\sigma_\epsilon^2}{\sigma_{T,\text{SIM1}}^2}
\]
Solving this equation resulted in a $\sigma_5^2$ of 0.01070956 for $\beta$-Cryptoxanthin, 0.004405707 for lycopene, and 0.02127483 for folate. Finally, at each iteration we simulate $\sigma_6^2$ and then calculate $\sigma_U^2 = \sigma_5^2 - \sigma_6^2$.

Lastly, we simulate the time-to-event from an exponential distribution with parameter $\lambda = \lambda_0 \exp (\beta_1 X + \beta_2 \text{AGE} + \beta_3 \text{BMI})$. The true regression parameters, $\beta_j$, $j = 1, 2, 3$, were based on a hypothesized strength of association between each exposure and diabetes that resulted in approximately 90% power $\beta_j$. In particular, we chose $\beta_1$ to be $\log(0.775) = -0.255$ for $\beta$-cryptoxanthin, $\log(0.529) = -0.637$ for Lycopene, and $\log(0.665) = -0.408$ for Folate, which correspond to hazard ratios for a 20% increase in true nutrient consumption of 0.954, 0.890, and 0.928, respectively. Further, we selected $\beta_2 = \log(0.9)$ and $\beta_3 = \log(0.75)$. To mimic diabetes incidence in the HCHS/SOL cohort, we simulate a censoring rate of approximately 85% by selecting a final censoring time of 60 for all simulations, which is meant to represent time in months, and fixing $\lambda_0$ at 650, 2000, and 1600 for $\beta$-Cryptoxanthin, lycopene, and folate, respectively.

Finally, we note that for simplicity of presentation, we assumed HCHS/SOL was a simple random sample of United States Hispanic/Latinos rather than a design-based sample.
Supplemental Materials

**Web Table 1.** Variables of interest in building prediction models for each nutrient biomarker. Here are the results and we will find out how long the margins

| Variables          | α-carotene | α-tocopherol | B-12  | β-carotene | β-cryptoxanthin | Folate | γ-tocopherol | Lycopene | Retinol | Zeaxanthin |
|--------------------|------------|--------------|-------|------------|----------------|--------|--------------|----------|---------|------------|
| Baseline Variables | ×          | ×            | ×     | ×          | ×              | ×      | ×            | ×        | ×       | ×          |
| Cholesterol Variables | ×        | ×            |       | ×          | ×              | ×      | ×            | ×        | ×       | ×          |
| Season             | ×          | ×            | ×     | ×          |                | ×      |              |          |         |            |
| Physical Activity  | ×          | ×            | ×     | ×          | ×              |        |              |          |         |            |

1. Baseline variables included in the prediction equations for all nutrients consist of age, gender, ethnicity, language preference, BMI, education, income, smoking, alcohol, diabetes, family history of diabetes, hypertension, high cholesterol, log total energy intake, a binary supplement use indicator, the associated self-reported nutrient and the corresponding measure of supplement use for each nutrient (if it exists).

2. Cholesterol variables consist of HDL-cholesterol (mg/dL), LDL-cholesterol (mg/dL), and lipid lowering drugs/Antihyperlipidemics (LLD).
Web Table 2. Nutrient biomarkers and corresponding self-reported nutrients for use in prediction models.

| Biomarker               | Associated Self-Reported Nutrients                                      |
|-------------------------|-------------------------------------------------------------------------|
| \( \alpha \)-carotene (µg/dL) | \( \alpha \)-carotene (provitamin A carotenoid) (mcg)                  |
| \( \alpha \)-tocopherol (mg/dL) | Total \( \alpha \)-tocopherol Equivalents (mg), \( \beta \)-carotene (provitamin A carotenoid) (mcg)\(^1\) |
| B-12 (pg/ml)             | Vitamin B-12 (cobalamin) (mcg)                                         |
| \( \beta \)-carotene (µg/dL) | \( \beta \)-carotene (provitamin A carotenoid) (mcg), Total \( \alpha \)-tocopherol Equivalents (mg)\(^1\) |
| \( \beta \)-cryptoxanthin (µg/dL) | \( \beta \)-cryptoxanthin (provitamin A carotenoid) (mcg)              |
| Retinyl Palmitate (µg/dL) | Retinol (mcg), \( \beta \)-carotene (provitamin A carotenoid) (mcg)\(^1\) |
| Folate (nmol/L)          | Total Folate (mcg)                                                     |
| \( \gamma \)-tocopherol (mg/dL) | \( \gamma \)-tocopherol (mg)                                         |
| Lycopene (µg/dL)         | Lycopene (mcg) (carotene)                                              |
| Retinol (µg/dL)          | Retinol (mcg), \( \beta \)-carotene (provitamin A carotenoid) (mcg)\(^1\) |
| Zeaxanthin (µg/dL)       | Lutein + Zeaxanthin (mcg) (xanthophyll or oxygenated carotenoids)      |

\(^1\) Previous publications show associations between plasma \( \alpha \)-tocopherol and self-reported dietary \( \beta \)-carotene; plasma \( \beta \)-carotene and self-reported dietary \( \alpha \)-tocopherol; plasma retinyl palmitate and self-reported dietary retinol and \( \beta \)-carotene; plasma retinol and self-reported dietary \( \beta \)-carotene.
Web Table 3. Variables chosen for the reduced prediction model for each nutrient by AIC in a stepwise selection algorithm, along with R² values for the full and reduced models.

| Variable                        | α-carotene | α-tocopherol | B-12 | β-carotene | β-cryptoxanthin | Folate | γ-tocopherol | Lycopene | Retinol | Zeaxanthin |
|---------------------------------|------------|--------------|------|------------|-----------------|--------|--------------|----------|---------|------------|
| Corresponding Self-Report       | ×          | ×            |      | ×          | ×                | ×      | ×            |          |         |            |
| Additional Self-Report           |            | ×            |      |            |                 |        |              |          |         |            |
| Age                             | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        |         |            |
| Gender                          | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        |         |            |
| Ethnicity                       | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        | ×       | ×          |
| Language Preference             | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        |         | ×          |
| BMI                             | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        |         | ×          |
| Income                          | ×          |              |      |            |                 |        |              |          |         |            |
| Education                       | ×          |              |      |            |                 |        |              |          |         |            |
| Smoking                         | ×          | ×            | ×    | ×          | ×                |        |              |          |         |            |
| Alcohol                         | ×          |              |      |            |                 |        |              |          | ×       | ×          |
| Diabetes                        | ×          |              |      |            |                 |        |              |          |         |            |
| Family History of Diabetes      | ×          |              |      |            |                 |        |              |          |         |            |
| High Cholesterol                | ×          |              |      |            |                 |        |              |          |         |            |
| Hypertension                    | ×          |              |      |            |                 |        |              |          |         |            |
| Log Total Energy Intake         | ×          |              |      |            |                 |        |              |          |         |            |
| Supplement Use                  | ×          | ×            | ×    | ×          | ×                | ×      | ×            | ×        |         | ×          |
| Supplement Intake               | ×          |              |      |            |                 |        |              |          |         |            |
| HDL-Cholesterol                 | ×          |              |      |            |                 |        |              |          |         |            |
| LDL-Cholesterol                 | ×          | ×            |      | ×          | ×                |        |              |          |         | ×          |
| Lipid Lowering Drugs            | ×          |              |      |            |                 |        |              |          |         |            |
| Season                          | ×          |              |      |            |                 |        |              |          |         |            |
| Physical Activity               | ×          |              |      |            |                 |        |              |          |         |            |
| **R² Full**                     | 0.4692     | 0.3486       | 0.1013| 0.4562     | 0.5035          | 0.1717| 0.2340       | 0.2196   | 0.1185  | 0.2728     |
| **R² Stepwise**                 | 0.4518     | 0.3334       | 0.0797| 0.4469     | 0.4856          | 0.1416| 0.2211       | 0.1794   | 0.1060  | 0.2410     |
**Web Table 4.** \( R^2 \) values based on the full prediction models, Prentice \( R^2 \) values, biomarker intraclass correlation coefficients (ICC), correlations between the biomarker and associated self-reported 2-day mean, and new ICC and \( R^2 \) measures that could exist if there were 2 and 4 repeat measures of each biomarker available on everyone in the SOLNAS substudy.

| Nutrient          | \( R^2 \) | Prentice \( R^2 \) | ICC  | Correlation SR and Biomarker (\( \rho \)) | ICC\(_{\text{new}}\) (2)\(^1\) | \( R^2_{\text{new}} \) (2)\(^2\) | ICC\(_{\text{new}}\) (4)\(^1\) | \( R^2_{\text{new}} \) (4)\(^2\) |
|-------------------|-----------|---------------------|------|------------------------------------------|-------------------------------|------------------------------|-------------------------------|------------------------------|
| \( \beta \)-cryptoxanthin | 0.5035    | 0.5662              | 0.8892 | 0.4090                                   | 0.9412                        | 0.5329                       | 0.9697                        | 0.5491                       |
| \( \alpha \)-carotene     | 0.4692    | 0.5792              | 0.8101 | 0.3321                                   | 0.8948                        | 0.5183                       | 0.9445                        | 0.5470                       |
| \( \beta \)-carotene     | 0.4562    | 0.5258              | 0.8676 | 0.2801                                   | 0.9267                        | 0.4872                       | 0.9619                        | 0.5058                       |
| \( \alpha \)-tocopherol   | 0.3486    | 0.5304              | 0.6571 | 0.0442                                   | 0.7928                        | 0.4205                       | 0.8844                        | 0.4691                       |
| Zeaxanthin           | 0.2728    | 0.3918              | 0.6961 | 0.2447                                   | 0.8206                        | 0.3215                       | 0.9015                        | 0.3532                       |
| \( \gamma \)-tocopherol | 0.2340    | 0.3776              | 0.6196 | 0.0216                                   | 0.7651                        | 0.2889                       | 0.8669                        | 0.3273                       |
| Lycopene             | 0.2196    | 0.3214              | 0.6833 | 0.1002                                   | 0.8109                        | 0.2607                       | 0.8956                        | 0.2879                       |
| Folate               | 0.1717    | 0.2916              | 0.5887 | 0.1211                                   | 0.7388                        | 0.2154                       | 0.8498                        | 0.2478                       |
| Retinol              | 0.1185    | 1.2871              | 0.0920 | 0.0236                                   | 0.1757                        | 0.2261                       | 0.2989                        | 0.3847                       |
| VB12                | 0.1013    | 0.1283              | 0.7891 | 0.0474                                   | 0.8787                        | 0.5606                       | 0.9354                        | 0.5968                       |

1. ICC\(_{\text{new}}\) (\( j \)) is calculated as \( \frac{\text{Var}(X)}{\text{Var}(X)+\text{Var}(U)/j} \) and is interpreted as the new ICC measure that would be available if we had \( j \) repeat biomarker measures available on everyone in the SOLNAS substudy. \( \text{Var}(X) \) is the variance of the true nutrient measure calculated as \( \text{Var}(X^{**}) \times \text{ICC} \), where \( \text{Var}(X^{**}) \) is the variance of the biomarker measure. \( \text{Var}(U) \) is the variance of the measurement error of the biomarker.

2. \( R^2_{\text{new}} \) (\( j \)) is calculated as \( R^2 \times \frac{\text{Var}(X^{**})}{\text{Var}(X^{**})+\text{Var}(U)/j} \) and is interpreted as the new \( R^2 \) measure that would be available if there were \( j \) repeat biomarker measures available on everyone.
Web Figure 1. Study timing and procedures in the SOL Nutritional Biomarkers and Study (SOLNAS).

SOL Parent Study
Baseline visit*:
Day 1: Anthropometry, physical activity & other questionnaires and Actical Accelerometry (day 1-7)
In person 24-hour recall
Day 5-90: 2nd 24-hour recall
*FPQ collected at Year One Follow-up call for Parent Study

Not eligible
N=476 (Main study)
(5-210 days after the parent study visit)
▷ Informed consent
▷ Weight
▷ Pre-DLW spot urine
▷ DLW dosing
▷ 2 post DLW spot urines
▷ Blood draw for pts 60+ yrs old
▷ In-person 24-hour diet recall
▷ Body image, Sed. Beh. Q
▷ 24-hour urine instructions

SOLNAS Visit 1
N=476
12 days after SOLNAS Visit 1 (Day 13)
▷ Collect 24 hour urine
▷ Weight
▷ Fasting blood draw
▷ Fasting urine
▷ One spot urine
▷ Indirect calorimetry
▷ Multi-cultural Q (Bronx site only N=119)

SOLNAS Visit 2
N=96 (Reliability sub-sample)
5-7 months post SOLNAS Visit 1
▷ Weight
▷ Pre-DLW spot urine
▷ DLW dosing
▷ 2 post DLW spot urines
▷ Blood draw for pts 60+ yrs old
▷ In-person 24-hour recall
▷ Body image assessment
▷ 24-hour urine instructions
▷ Physical activity-Sed. Beh Q
▷ Actical & questionnaires

SOLNAS Visit 3
N=96 (Rel. subsample)
2nd 24-hour diet recall:
5-90 days post Visit #3

SOLNAS Visit 4
N=96 (reliability subsample)
12 days after SOLNAS Visit 3
▷ Collect 24-hour urine
▷ Weight
▷ Fasting blood draw
▷ Fasting urine
▷ Collect Actical data
▷ One spot urine
▷ Indirect calorimetry
▷ Multi-cultural FFQ (Bronx site only N=119)

For San Diego site only, SOLNAS Visit 1 window could be up to 12 months after parent study visit.
References

1. Harrell Jr FE. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer; 2015 Aug 14.
2. Frank E Harrell Jr, with contributions from Charles Dupont and many others. (2020). Hmisc: Harrell Miscellaneous. R package version 4.4-0. https://CRAN.R-project.org/package=Hmisc