

UC Davis

UC Davis Previously Published Works

Title

Effect of Manual Data Cleaning on Nutrient Intakes Using the Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24)

Permalink

<https://escholarship.org/uc/item/86q1j1mg>

Journal

Current Developments in Nutrition, 5(3)

ISSN

2475-2991

Authors

Bouزيد, Yasmine Y
Arsenault, Joanne E
Bonnell, Ellen L
[et al.](#)

Publication Date

2021-03-01

DOI

10.1093/cdn/nzab005


Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nc-nd/4.0/>

Peer reviewed



Effect of Manual Data Cleaning on Nutrient Intakes Using the Automated Self-Administered 24-Hour Dietary Assessment Tool (ASA24)

Yasmine Y Bouzid,^{1,2} Joanne E Arsenault,² Ellen L Bonnel,^{1,2} Eduardo Cervantes,^{1,2} Annie Kan,^{1,2} Nancy L Keim,^{1,2} Danielle G Lemay,^{1,2} and Charles B Stephensen^{1,2} 

¹USDA Agricultural Research Service Western Human Nutrition Research Center, Davis, CA, USA and ²Department of Nutrition, University of California, Davis, Davis, CA, USA

ABSTRACT

Background: Automated dietary assessment tools such as ASA24[®] are useful for collecting 24-hour recall data in large-scale studies. Modifications made during manual data cleaning may affect nutrient intakes.

Objectives: We evaluated the effects of modifications made during manual data cleaning on nutrient intakes of interest: energy, carbohydrate, total fat, protein, and fiber.

Methods: Differences in mean intake before and after data cleaning modifications for all recalls and average intakes per subject were analyzed by paired t-tests. The Chi-squared test was used to determine whether unsupervised recalls had more open-ended text responses that required modification than supervised recalls. We characterized food types of text response modifications. Correlations between predictive energy requirements, measured total energy expenditure (TEE), and mean energy intake from raw and modified data were examined.

Results: After excluding 11 recalls with invalidating technical errors, 1499 valid recalls completed by 393 subjects were included in this analysis. We found significant differences before and after modifications for energy, carbohydrate, total fat, and protein intakes for all recalls ($P < 0.05$). Limiting to modified recalls, there were significant differences for all nutrients of interest, including fiber ($P < 0.02$). There was not a significantly greater proportion of text responses requiring modification for home compared with supervised recalls ($P = 0.271$). Predicted energy requirements correlated highly with TEE. There was no significant difference in correlation of mean energy intake with TEE for modified compared with raw data. Mean intake for individual subjects was significantly different for energy, protein, and fat intakes following cleaning modifications ($P < 0.001$).

Conclusions: Manual modifications can change mean nutrient intakes for an entire cohort and individuals. However, modifications did not significantly affect the correlation of energy intake with predictive requirements and measured expenditure. Investigators can consider their research question and nutrients of interest when deciding to make cleaning modifications. *Curr Dev Nutr* 2021;5:nzab005.

Keywords: 24-hour recall, ASA24, dietary assessment, data cleaning, total energy expenditure

Published by Oxford University Press on behalf of the American Society for Nutrition 2021. This work is written by (a) US Government employee(s) and is in the public domain in the US.

Manuscript received December 17, 2020. Initial review completed January 15, 2021. Revision accepted January 26, 2021. Published online February 2, 2021.

This study was funded by USDA Intramural Projects 2032-51530-022-00D, 2032-51530-025-00D, and 2032-51530-026-00-D. The USDA is an equal opportunity employer and provider. Additional support was provided by the National Center for Advancing Translational Sciences, NIH, through grant number UL1 TR001860. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author disclosures: The authors report no conflicts of interest.

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

Address correspondence to CBS (e-mail: charles.stephensen@usda.gov).

Abbreviations used: AMPM, Automated Multi-Pass Method; ASA24, Automated Self-Administered 24-Hour Dietary Assessment Tool; FNDDS, Food and Nutrient Database for Dietary Studies; NCI, National Cancer Institute; PA, physical activity; RMR, resting metabolic rate; TEE, total energy expenditure; TEF, thermic effect of food.

Introduction

The manual collection of 24-hour dietary recalls in large-scale studies can be time-prohibitive as highly trained personnel are needed to conduct interviews and process the data (1). The development of web-based tools like the Automated Self-Administered 24-Hour (ASA24[®]) Dietary Assessment Tool have expanded the potential for collecting 24-hour recalls in large-scale studies, with little burden on dietary experts to collect, review, and code data (2). However, given the inherent complexity of obtaining accurate dietary data (3), the performance of automated

tools must be repeatedly assessed to identify systematic and user errors and their impact on measurement quality. For example, users can fill in open-ended text responses to prompts in ASA24 which are then automatically coded to foods in the Food and Nutrient Database for Dietary Studies (FNDDS). Previous work by Zimmerman et al. evaluated the impact of correcting suboptimal matches for "Other, specify" and "Unfound food" open-ended text responses in the ASA24-2011 version (4). They found significant differences between intakes of energy, protein, fat, and carbohydrate before and after data cleaning ($P < 0.05$). Although the scale of these differences was not considered large enough

to warrant broad recommendations to inspect and correct open-ended text responses, suboptimal matches may have a more profound effect when investigating individual dietary intake or certain nutrients and food groups affected by automated coding that was suboptimal.

The ASA24 system was validated in a study comparing reported intake to observed true intake that identified some common errors in reporting (5). Manual data cleaning was conducted by inspecting text responses and recoding foods where default matches were suboptimal. Sources of error identified in this study include subjects being unable to find a food through the search tool and writing in an “Unfound food.” When automatically coded, sometimes this would result in suboptimal matches. Researchers also observed a tendency for subjects to select the first food on the search list options instead of searching for and selecting the closest match. Corrections were also made for “Known Issues,” which are provided by ASA24 for researchers. In the current study, we characterize the food type and frequency of modifications made for text response mismatches.

Previous studies have investigated how assistance influences participants’ interaction with the ASA24 system (6–8). Factors that contribute to misreporting when assistance is provided by trained personnel include concerns for maintaining social desirability (8) and in a study assessing the usability of ASA24, inability to find search items, knowingly reporting incorrect information, misinterpreting questions, and “misclicks” were issues identified (7). These issues can affect the quality of dietary data collected and guidance by trained personnel may decrease the number of errors that occur relative to recalls completed independently. We are interested in how recall training and assistance may influence the number of modifications needed for data cleaning. Thus, an outstanding question is whether subjects write in more text responses that lead to suboptimal matches while completing recalls independently. In the current study, we assess proportions of modifications made to supervised training recalls compared with unsupervised home recalls.

Inquiries into precision nutrition will necessitate assessment of individual intake within sufficiently powered, large-scale studies, with multiple recalls per individual, using tools with validated assessment methods and databases. With low researcher burden and Automated Multi-Pass Method (AMPM) design (9), ASA24 is a logical choice for data collection that may be used for individual intake assessment. A recent review evaluating dietary data collection toolkits stated that administration of multiple 24-hour recalls in a cohort is sufficient to describe usual and acute intake distributions (10). However, this requires an adjustment for measurement error through statistical modeling such as the National Cancer Institute (NCI) method to examine diet-disease relations (11). For the current analysis, we compare nutrient intakes without usual intake adjustment for a more direct comparison of the magnitude of differences and the method cannot be applied for a secondary analysis comparing the first recall that was supervised to subsequent unsupervised recalls. In addition, we assess the effect of data cleaning on dietary assessment of individual participants by examining the correlation between reported energy intake with estimated energy requirements and measured total energy expenditure (TEE).

The association between errors while using ASA24 and participant characteristics, such as age, sex, and BMI is not fully known. A study examining ASA24 versus the previous interviewer-assisted AMPM method found participants expressed preference for ASA24 across age groups and educational levels (8). In the present study, a var-

ied population of participants were recruited by stratifying for sex, age, and BMI. Given that ASA24 is a preferred tool that decreases participant burden for diverse groups, we sought to explore associations between data cleaning modifications and recruitment group characteristics on nutrients of interest: energy, carbohydrate, fat, protein, and fiber intake.

In the current study, participants completed up to 4 recalls in the ASA24-2014 and 2016 versions and data was cleaned based on NCI recommendations and by correcting nutrient and food group information for open-ended text responses that produced suboptimal matches. Our goal was to determine whether nutrient intakes differed before and after modifying recalls collected from a cross-sectional nutritional phenotyping study. We provide qualitative descriptions of food and supplement text responses that produced suboptimal matches and corrections made. Estimated energy requirements and TEE were calculated and compared with reported energy intake to assess correlation of raw and modified data to an objective measure.

Methods

Participant population

Participants were recruited in the USDA Nutritional Phenotyping Study (12). The study is registered at clinicaltrials.gov as NCT02367287. Participants were recruited based on an 18-bin sampling scheme stratified by sex, age, and BMI. There were 9 bins for each sex, male and female. Within these, age ranges were divided into 3 categories (18.00–33.99 y, 34.00–49.99 y, 50.00–65.99 y). Then, BMI range was divided into 3 categories according to WHO international classifications (“normal”: 18.50–24.99 kg/m², “overweight”: 25.00–29.99, and “obese”: 30.00–45.00) (12). Study enrollment took place from May 2015 to July 2019. Specifics of the recruitment process are presented in a CONSORT diagram (**Supplemental Figure 1**). The study was reviewed and approved by the University of California, Davis, Institutional Review Board. All participants provided written informed consent and received monetary compensation for their participation. Data were stored using the Research Electronic Data Capture (REDCap) application hosted by the University of California Davis Health System Clinical and Translational Science Center.

Dietary data collection

The data collection methodology of ASA24 recalls in this study was described previously (12). Briefly, trained personnel instructed participants on using the ASA24 system by guiding them on navigating between pages and accurately searching for foods while they completed a supervised recall at the first study visit. Personnel directed participants on how to report a meal, search for foods in the database, and add details for this supervised recall. Measuring cups to help with portion estimation were presented as supplementary aids to the pictures generated by ASA24. During the following 10–14 days, participants were prompted to complete 3 unannounced recalls using ASA24 at home, reporting all food, drink, and supplements consumed from 12:00am to 11:59pm of the previous day.

Dietary data cleaning

Based on ASA24’s recommended guidelines for data cleaning (13), we investigated foods and supplements that ASA24 marked as “Data

Missing” in the ASA24-2014/2016 output data files with probe questions and answers by recall (MS/Responses) and nutrient data by food and beverage item (INFMYPHEI/Items). We categorized the reasons for missing data which included nutrient data missing for foods identified in the ASA24 database, nutrient data missing for “Other, specify” and “Unfound food” entries, and nutrient data missing for supplements. Some food items were flagged for review because subjects responded “Yes” to the ASA24 prompt “Anything added?” to the food, then responded “Don’t know” or “Nothing added” to the follow-up prompt, which created a blank entry with no nutrient data. There were a few instances where participants reported a food item and amount, but no nutrient data was autocoded. These were corrected during data cleaning. Otherwise, most of the recalls flagged by ASA24 with “Data Missing” did not require modification.

A registered dietitian (JEA) reviewed all “Other, specify” and “Unfound food” open-ended text responses and evaluated the appropriateness of the automated coding match. Modifications were made to nutrient data of foods and supplements where automated coding matches were judged as suboptimal based on text responses. We prioritized modifying suboptimal matches that most likely impacted macronutrient and fiber intake. For example, these included almond milk that was not available in the 2014 database, salsa or pesto coded as mayonnaise, soy products coded as meat, burritos coded with meat when not indicated, brown rice coded as white rice, and fish oil coded as the default supplement. SAS code provided by ASA24 (13) was used to convert ASA24-2014 My Pyramid Equivalents Database (MPED 2.0) food group intakes to ASA24-2016 Food Patterns Equivalents Database (FPED 2011–2012) for consistency.

We examined “Known Issues” reported by ASA24 (14) and no changes needed to be made from these. Outliers for portions and nutrients were reviewed according to ASA24 cleaning guidelines (13) and modifications were made after consensus by the study team. Some recalls were judged to be invalid and excluded from analysis if the participant did not complete all prompting questions or staff assessed the recalls were not completed properly.

Anthropometrics, TEE measurement and calculation

Height was measured in duplicate to the nearest 0.1 cm using a permanently mounted stadiometer (Ayrton). Weight was measured in duplicate to the nearest 0.1 kg using an electronic scale (Tanita® BWB-627A). BMI was calculated and expressed as kg/m². Energy requirements were estimated using Harris Benedict, Mifflin St. Jeor, and Cunningham equations. TEE was calculated using measures of activity and resting metabolic rate (RMR). Participants were instructed to wear a Respironics® Actical™ accelerometer around their waist for 7 days between the first and second study visit to approximately measure energy expended from physical activity (PA). Fasted RMR was measured at the second study visit using indirect calorimetry carts (TrueOne 2400, Parvo Medics) (12). Thermic effect of food (TEF) was calculated as 10% of measured RMR. TEE was calculated as the sum of RMR, TEF, and PA.

Statistical analysis

To evaluate differences in energy, carbohydrate, fat, protein, and fiber intakes between the raw data output from ASA24 and modified data, nutrients were first normalized using Tukey’s Ladder of Powers trans-

formation and paired *t*-tests were performed between individual recalls and mean intake per subject before and after modification. The percent change between raw and modified nutrients was calculated to determine the proportion of modified recalls changed by >10% due to modification, as examined by Zimmerman et al. (4). Pearson’s chi-squared test was used to determine whether significantly more modifications were made to recalls conducted independently at home than the supervised recall at the first study visit. Ten recalls with the greatest percent increase and greatest percent decrease for each nutrient were identified to describe types of modifications that resulted in these changes. Group analysis was conducted using mixed linear regression models to determine the effect of modifications, age, sex, and BMI on nutrient intakes and assess interactions of these covariates with subject as a random variable. Tukey’s Ladder of Powers was used to transform nutrient distributions prior to mixed model analysis. Estimated energy requirements were compared with measured TEE by the Pearson correlation procedure. Estimated energy requirements were also compared with raw and modified energy intake estimates. Mean energy intake was compared with measured TEE using the Pearson correlation for subjects with 2 or more complete recalls. Tukey’s Ladder of Powers was used to transform the TEE distribution. *P* values < 0.05 were considered statistically significant. All statistical analyses were conducted using R software (version 3.6.0).

Results

Participant characteristics

Table 1 shows outcomes of the recruitment strategy outlined previously (12). Participants living in the greater Sacramento region were recruited. The 393 participants include all who completed the first study visit. These participants would have at least completed the supervised recall on-site with study personnel.

Effect of cleaning on individual recalls

After removing 11 recalls that were incomplete or with technical errors that made the recall unreliable, 1499 valid recalls completed by 393 participants were used in this analysis. The single supervised recall (which was the first of 4 possible recalls per participant) was completed by 393 subjects, 8 subjects completed just 2 recalls (supervised plus 1 at-home recall), 35 completed 3 recalls (supervised plus 2 at-home recalls), and 343 completed all 4 recalls. Due to unknown circumstances, data from 2 supervised recalls were not saved automatically at the time of the study visit and were thus not available for this analysis. A total of 233 recalls (16%) were modified during the data cleaning process (Table 2); 166 recalls (11%) were reviewed because ASA24 flagged them with “Data Missing” and 917 recalls (61%) had open-ended text responses that were reviewed.

Figure 1 shows plots of the nutrient intakes for raw and modified data for the 233 modified recalls. There were 151 participants with at least 1 recall modified (38% of all participants). When limiting to recalls modified during cleaning (*n* = 233), significant differences were found for all nutrients (*P* < 0.05) (Table 3). In this case, mean fiber intake increased significantly after modifications. Using all 1499 recalls, mean intakes of energy, carbohydrate, total fat, and protein were

TABLE 1 Characteristics of participants based on study design bins on sex, age, and BMI ($n = 393$)

Sex	Age (y)	BMI	Bin number	Age	BMI
				(mean \pm SD)	(mean \pm SD)
Male ($n = 184$)	18–33 ($n = 66$)	<25.0	1 ($n = 28$)	24.39 \pm 3.65	22.51 \pm 1.95
		25.0–29.9	2 ($n = 24$)	23.12 \pm 2.52	27.20 \pm 1.41
		≥ 30.0	3 ($n = 14$)	26.79 \pm 3.45	33.57 \pm 1.96
	34–49 ($n = 63$)	<25.0	4 ($n = 24$)	39.62 \pm 5.47	23.43 \pm 1.24
		25.0–29.9	5 ($n = 22$)	39.64 \pm 4.88	27.51 \pm 1.46
		≥ 30.0	6 ($n = 17$)	43.35 \pm 4.42	35.05 \pm 3.95
	50–65 ($n = 55$)	<25.0	7 ($n = 19$)	58.26 \pm 4.13	22.71 \pm 1.46
		25.0–29.9	8 ($n = 27$)	56.30 \pm 5.53	27.36 \pm 1.53
		≥ 30.0	9 ($n = 9$)	56.44 \pm 4.10	33.91 \pm 1.89
Female ($n = 209$)	18–33 ($n = 74$)	<25.0	10 ($n = 26$)	24.73 \pm 4.25	22.07 \pm 1.55
		25.0–29.9	11 ($n = 26$)	26.04 \pm 3.84	27.99 \pm 1.39
		≥ 30.0	12 ($n = 22$)	24.45 \pm 3.94	34.44 \pm 3.33
	34–49 ($n = 67$)	<25.0	13 ($n = 25$)	41.56 \pm 4.36	22.52 \pm 1.90
		25.0–29.9	14 ($n = 19$)	42.21 \pm 4.88	27.33 \pm 1.48
		≥ 30.0	15 ($n = 23$)	41.13 \pm 5.00	34.59 \pm 3.55
	50–65 ($n = 68$)	<25.0	16 ($n = 23$)	55.70 \pm 4.47	22.61 \pm 1.25
		25.0–29.9	17 ($n = 26$)	57.46 \pm 4.22	27.11 \pm 1.47
		≥ 30.0	18 ($n = 19$)	56.26 \pm 4.8	33.98 \pm 2.98

different after data cleaning modifications compared with before ($P < 0.05$). Modifications decreased the mean intakes for these 4 nutrients.

Table 4 shows how many modified recalls resulted in changes to nutrient intakes greater than 10%.

Supervised versus unsupervised recall modifications

When considering all of the modified recalls ($n = 233$), there was a greater proportion of unsupervised home (16%) than supervised (14%) recalls modified, but not significantly ($P = 0.271$). A total of 179 of 1108 unsupervised recalls and 54 of 391 supervised recalls were modified.

Characterization of the types of modifications

For all open-ended text responses, 92% of “Other, specify” and 73% of “Unfound food” responses did not require modification. Modifications were made to 268 items judged to have suboptimal automated coding matches for write-in text responses during review by the registered dietitian (**Table 5**). Of these text response modifications, 157 (59%) were “Other, specify” and 111 (41%) were “Unfound food” write-ins. Non-text modifications were made for 27 items, for a total of 295 items modified during data cleaning. We reviewed the modifications made for suboptimal matches of “Unfound food” and “Other, specify” open-ended text responses and qualitatively categorized the food types that required modifications (**Figure 2**). Examples of frequent categories are noted, and further examples are in **Supplemental Table 1**.

Most modifications made were in the “meat_cheese_egg_veg” category (**Figure 2**). This describes text responses that were miscoded as to

the form of meat, cheese, egg, or vegetables. For example, a salad was reported with tofu as an “Other, specify” ingredient, and coded as cheese. This was changed to soybean curd. In another case, a sandwich with tomatoes as an “Other, specify” food was coded as eggs and corrected to tomatoes. The next most frequent category for modifications was for supplements, with the largest proportion being reported as “Unfound food” (**Figure 2**). When a subject reported a supplement by selecting “Match not found”, the default supplement coded was fish oil. However, sometimes a supplement write-in was coded as a food item (examples include peanut butter, rice, soft drink). Within the beverage category, most modifications were attributed to milk products, as denoted by a separate “beverage_milk” category (**Figure 2**). This is largely due to ASA24-2014 lacking almond milk and an a priori decision to correct write-ins for consistency with the 2016 database. In a few cases, subjects tried to report kombucha or kefir as “Unfound foods,” and automated coding produced catsup. In one case, a write-in response for “Unfound food” of “water, plain spring” was coded as chocolate syrup.

We examined modification types for 10 recalls with the greatest percent decrease and increase for nutrient estimates (**Supplemental Table 2**). Most modifications that resulted in recalls with the greatest percent decrease in energy intakes were for “Outlier corrections” (8 out of 10 recalls). Modifications that resulted in the greatest percent increase for energy were due to suboptimal text response matches, a majority from “Unfound food” responses (13/19 items in 10 recalls). The recall with the largest percent increase in energy (54%) had 3 “Unfound food” text responses that led to suboptimal matches,

TABLE 2 Number of recalls modified and items per recall

	Total valid recalls	Recalls modified	Number of items modified = 1	Number of items modified = 2	Number of items modified = 3	Number of items modified = 4	Number of items modified = 5
	1499	233	189	34	3	6	1
Percentage of total		16%	13%	2%	0.2%	0.4%	0.06%
Percentage of modified			81%	15%	1%	3%	0.4%

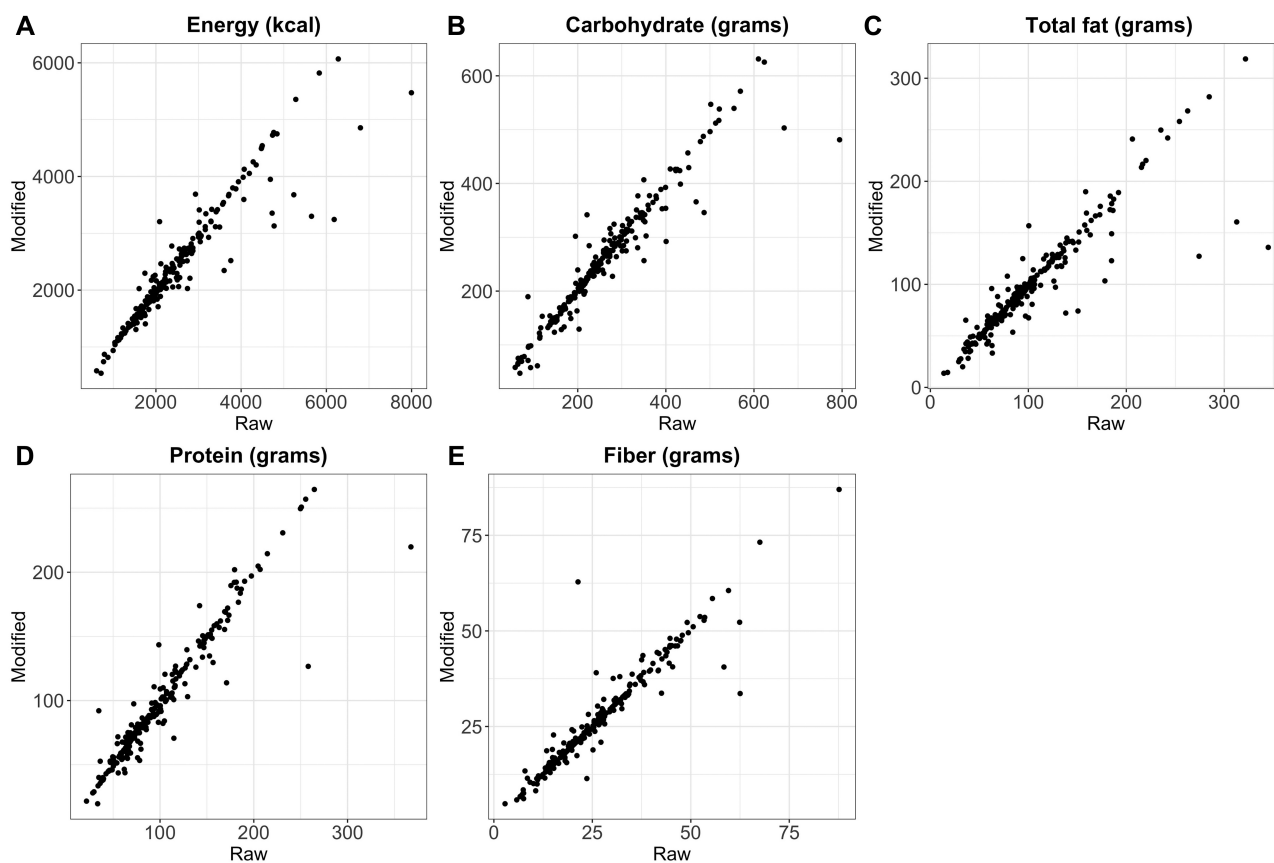


FIGURE 1 Plots of nutrient intakes from food and supplements before and after cleaning modifications for modified recalls ($n = 233$)

categorized as “beverage,” “meat_cheese_egg_veg,” and “snack.” Of the 10 modified recalls that resulted in the greatest percent decrease in carbohydrate intake, there were 8 “Unfound food” and 4 “Other, specify” text response modifications, 3 “Outlier corrections,” and 1 “Database error.” The most frequent text response category was “beverage” re-

ported 4 times as an “Unfound food.” For the 10 recalls with the greatest percent increase in carbohydrate, there were 12 “Unfound food” text responses and 2 “Outlier corrections.” The recall with the greatest percent increase (118%) resulted from modification to a “beverage” item and “supplement” was the most common text response with 4

TABLE 3 Differences in nutrient intakes from food and supplement data before and after cleaning modifications

	Nutrient	Before modifying (untransformed mean \pm SD)	After modifying (untransformed mean \pm SD)	Difference (after minus before)	P value
All recalls ($n = 1499$)	Energy, kcal	2214.9 \pm 955.6	2199.4 \pm 923.5	-15.5	<0.0001
	Carbohydrate, g	243.3 \pm 117.6	242.8 \pm 116.4	-0.5	<0.0001
	Total fat, g	92.5 \pm 47.6	91.8 \pm 46.6	-0.7	<0.0001
	Protein, g	94.6 \pm 49.0	94.3 \pm 48.4	-0.3	0.041
	Fiber, g	24.0 \pm 14.1	24.1 \pm 14.1	0.1	0.16
All recalls from participants with ≥ 1 modified ($n = 593$)	Energy, kcal	2366.4 \pm 1015.1	2327.3 \pm 942.8	-39.1	<0.0001
	Carbohydrate, g	255.7 \pm 116.3	254.4 \pm 113.3	-1.3	<0.0001
	Total fat, g	100.6 \pm 51.6	98.8 \pm 49.4	-1.8	<0.0001
	Protein, g	98.6 \pm 50.6	97.6 \pm 49.2	-1.0	0.083
	Fiber, g	25.5 \pm 12.8	25.6 \pm 12.7	0.1	0.374
Modified recalls ($n = 233$)	Energy, kcal	2440.1 \pm 1116.2	2340.7 \pm 945.0	-99.4	<0.0001
	Carbohydrate, g	259.0 \pm 116.1	255.7 \pm 108.3	-3.3	<0.0001
	Total fat, g	102.1 \pm 54.7	97.4 \pm 49.3	-4.7	<0.0001
	Protein, g	100.8 \pm 51.3	98.5 \pm 47.9	-2.3	0.004
	Fiber, g	26.3 \pm 12.8	26.6 \pm 12.8	0.3	0.016

P values from paired t-test before and after cleaning.

TABLE 4 Percentage of modified recalls ($n = 233$) with changes in nutrient intake $>10\%$

Nutrient	Percentage of 233 recalls that were modified during data cleaning		
	Changed by $>10\%$	Decreased by $>10\%$	Increased by $>10\%$
Energy	15.0	10.7	4.3
Carbohydrate	16.8	9.9	6.9
Total fat	23.6	15.4	8.2
Protein	18.9	10.7	8.2
Fiber	16.3	6.0	10.3

modifications made to write-ins. The greatest percent decrease in total fat intakes from modifications were due to “Outlier corrections” and “Other, specify” text responses (4 and 9 modifications, respectively, in 10 recalls). Recalls with the greatest percent increase in total fat intakes had text response modifications (8 “Other, specify” and 10 “Unfound food”). Two recalls with the greatest percent increase in total fat had an “Unfound food” “meat_cheese_egg_veg” modification (80.5% increase) and “beverage,” “meat_cheese_egg_veg,” and “snack” “Unfound food” entries (56.1% increase). The greatest decreases in protein intakes across 10 recalls were from 9 “Other, specify” and 2 “Unfound food” text responses and 3 “Outlier corrections.” The recall with the largest percent decrease (50.9%) was an “Outlier correction” and the recall with the second largest decrease (41.5%) had 2 “Other, specify” modifications categorized as “beverage_milk.” Recalls with the greatest percent increase in protein intakes had all text response modifications (5 “Other, specify” and 10 “Unfound food” across 10 recalls) with “meat_cheese_egg_veg” categorized from “Other, specify” and “Unfound food” responses for the greatest 2 recalls (165% and 45.5% increase, respectively). For fiber intakes, the 10 recalls with the greatest percent decreases had 8 “Other, specify” and 2 “Unfound food” text response modifications, 2 “Outlier corrections,” and 1 “Database error” modification. The greatest percent increases for fiber intake were due to 7 “Other, specify” and 6 “Unfound food” text response modifications, and 1 “Outlier correction” resulting in the recall with the greatest percent increase (194%).

Effect of cleaning on dietary assessment for individual participants

For participants with at least 2 valid recalls ($n = 385$), mean nutrient intakes from food and supplements had significant differences in energy, fat, and protein comparing all data before and after modification

TABLE 5 Number of items modified and reasons for modification

Total modifications	Text response modifications		Non-text response modifications			
	Other, specify	Unfound food	System glitch	Database error	Outlier correction	Omission
295	157	111	7	10	9	1
Percentage of total	53%	38%	2.4%	3.4%	3.0%	0.3%

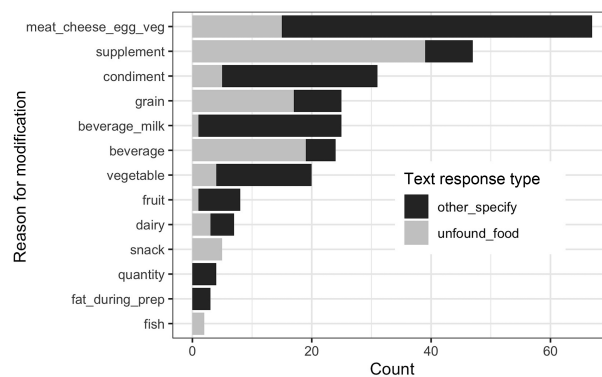
Descriptions and examples of non-text response modification categories

System glitch: An item and portion size were reported using the ASA24 search tool, but no nutrient data was on output files. Example: cooked carrots were reported on Responses file, but no nutrient data was on Items file.

Database error: Implausible nutrient amounts were assumed in FNDDS 2011–12 and corrected. Example: tea, hibiscus had 20 mg iron per cup in FNDDS.

Outlier correction: ASA24 provides recommendations to evaluate recall outliers for energy, beverage amounts, portion sizes etc. Corrections were made based on professional judgment for outliers. Example: 48 ounces (four 12 oz glasses) of whiskey were reported, assumed to mean 4 shots and amount corrected.

Omission: Participant omitted item from recall and later reported to study staff. Example: 56 ounces (seven 8 oz glasses) of water added to recall.

**FIGURE 2** Frequency of reasons for text response modifications were classified by food type, whereby inappropriately coded items were changed to a food item that more closely matched the food reported by the respondent. Other reasons for modifications included the quantity of food was inappropriately coded and fat during preparation where a subject did not report oil that was coded.

($P < 0.001$) (Table 6). For participants with at least 1 modified recall ($n = 151$), significant differences for mean intakes of all 5 nutrients were found ($P < 0.001$). Participants with measured TEE ($n = 352$) had a slightly higher, although not significantly, correlation with mean energy intake for modified ($r = 0.430$) versus raw data ($r = 0.427$) (Table 7). This was also observed when limiting to participants with at least 1 modified recall ($n = 141$) for TEE correlated with mean energy intake from modified ($r = 0.430$) versus raw ($r = 0.426$) data. Estimated energy requirements calculated using predictive equations (Harris-Benedict, Mifflin St. Jeor, and Cunningham) were highly correlated with measured TEE for all participants.

Effect of cleaning on group intakes

For all 1499 valid recalls, linear mixed models were created to discern the effect of data cleaning on transformed nutrient intakes when accounting for age, sex, and BMI (Table 8). There were no significant interactions between modification status, age, sex, or BMI. Modifications did not have a significant effect on nutrient intakes when controlling for other factors that may be associated with nutrient intake. About half of variance was attributed to the individual (48–55%). Males had higher nutrient intakes compared with females based on this model

TABLE 6 Differences in mean nutrient intakes per individual from food and supplement data before and after cleaning modifications

	Nutrient	Before modifying (untransformed mean \pm SD)	After modifying (untransformed mean \pm SD)	Difference (after minus before)	P value
Participants with ≥ 2 recalls (n = 385)	Energy, kcal	2211.3 \pm 752.2	2196.3 \pm 732.5	-15	<0.0001
	Carbohydrate, g	242.8 \pm 95.8	242.4 \pm 95.2	-0.4	0.228
	Total fat, g	92.3 \pm 36.0	91.6 \pm 35.3	-0.7	<0.0001
	Protein, g	94.9 \pm 38.6	94.5 \pm 38.3	-0.4	<0.0001
Participants with ≥ 1 modified recall (n = 151)	Fiber, g	24.0 \pm 11.8	24.0 \pm 11.8	0	0.265
	Energy, kcal	2361.5 \pm 785.8	2323.4 \pm 743.9	-38.1	<0.0001
	Carbohydrate, g	255.4 \pm 93.4	254.2 \pm 91.9	-1.2	<0.0001
	Total fat, g	100.3 \pm 38.9	98.5 \pm 37.5	-1.8	<0.0001
	Protein, g	98.4 \pm 40.3	97.5 \pm 39.8	-0.9	<0.0001
	Fiber, g	25.4 \pm 9.9	25.5 \pm 10.0	0.1	<0.0001

P values from paired t-test before and after cleaning.

($P < 0.001$). Age also had a significant effect on protein intakes. Age and BMI had a significant effect for fiber, with fiber intake being higher in older participants and lower in participants with a higher BMI.

Discussion

This study provides quantitative and qualitative descriptions of the effects of dietary data cleaning that extends the work by Zimmerman et al. (4). Literature from studies using ASA24 do not usually report the effects on nutrient intakes of following ASA24's recommended dietary data cleaning protocol. Zimmerman et al. provided some examples of types of foods reported as "Other, specify" or "Unfound food" that required modification, but no systematic categorization of food types. Differences between raw and modified data for nutrient intakes for all recalls were reported, but these were not evaluated for each individual, as in the present study. We also investigated the effect of data modifications on dietary assessment accuracy using TEE as a physiological parameter. The current study provides a descriptive account of conducting data cleaning and modifications on a large dietary dataset from a nutritional phenotyping study, in consideration of group and individual dietary assessment.

Although a low proportion of all valid recalls were modified (16%), we found significant differences before and after data modifications for

energy and macronutrients (Tables 2 and 3). This is consistent with findings by Zimmerman et al., along with our finding that mean nutrient intakes were lower for modified than raw data. The magnitude of mean nutrient differences before and after modification was similar to findings in Zimmerman et al. for energy, carbohydrate, total fat, and protein. Although statistically significant, mean differences were small for all recalls, but larger for recalls that were modified. In qualitatively characterizing the types of text response modifications, we provide some insight into the differences between nutrient intakes for individual recalls. The most frequent category was "meat_cheese_egg_veg" which includes write-ins for vegetables such as tomatoes being coded as eggs (Figure 2). This may explain why the mean protein intake reported was significantly higher for raw compared with modified recall data (Table 3). There was high incidence in modifications for animal products being miscoded for plant foods in the "meat_cheese_egg_veg" category, potentially also underlying energy and fat intakes being higher for raw compared with modified recalls. "Outlier corrections" accounted for most of the modifications resulting in recalls with the greatest percent decrease in energy intake. There were also "Outlier corrections" in the 10 recalls with the greatest percent decreases in total fat and protein intakes.

There was no difference in the proportion of modifications made to supervised versus unsupervised recalls, suggesting that assistance does not necessarily prevent errors related to automated coding suboptimal

TABLE 7 The Pearson correlation between energy intake of raw and modified data, predictive energy requirements, and measured TEE, and for all participants with TEE and predictive equation data (A) and participants with at least 1 modified recall (B)

A)	Harris-Benedict (n = 345)	Mifflin St. Jeor (n = 345)	Cunningham (n = 345)	Measured TEE (n = 352)
Energy estimate before modifying	0.498	0.519	0.571	0.427
Energy estimate after modifying	0.501	0.521	0.576	0.430
Measured TEE	0.746	0.743	0.727	
B)	Harris-Benedict (n = 137)	Mifflin St. Jeor (n = 137)	Cunningham (n = 137)	Measured TEE (n = 141)
Energy estimate before modifying	0.539	0.558	0.559	0.426
Energy estimate after modifying	0.536	0.556	0.561	0.430
Measured TEE	0.692	0.694	0.687	

TEE, total energy expenditure.

TABLE 8 Summary of mixed effects models for nutrient intakes from food and supplements ($n = 1499$ recalls)

Predictor	Energy (kcal)		Carbohydrate (g)		Total fat (g)		Protein (g)		Fiber (g)	
	Estimates	P	Estimates	P	Estimates	P	Estimates	P	Estimates	P
Modified vs. raw	0.05 (-0.13-0.22)	0.607	0.01 (-0.11-0.12)	0.889	0.01 (-0.04-0.07)	0.619	0.00 (-0.03-0.04)	0.818	-0.00 (-0.03-0.03)	0.875
Age, y	-0.00 (-0.02-0.02)	0.853	0.01 (-0.01-0.02)	0.452	-0.00 (-0.01-0.00)	0.728	-0.00 (-0.01-0.00)	0.017	0.01 (0.00-0.01)	0.003
Male	2.71 (2.20-3.22)	<0.001	1.57 (1.18-1.96)	<0.001	0.59 (0.44-0.73)	<0.001	0.52 (0.42-0.61)	<0.001	0.30 (0.19-0.41)	<0.001
BMI, kg/m ²	0.01 (-0.04-0.06)	0.750	-0.04 (-0.08-0.00)	0.059	0.01 (-0.00-0.03)	0.077	0.01 (-0.00-0.02)	0.095	-0.03 (-0.04-0.02)	<0.001

The nutrient intakes were transformed by Tukey's Ladder of Powers and the transformations were as follows: energy^{0.4}, carbohydrate^{0.45}, total fat^{0.375}, protein^{0.325}, and fiber^{0.375}.

matches from text responses, which accounted for most of the modifications (Table 5). This is consistent with studies examining the accuracy and usability of ASA24 recalls in low-income adult populations. In a group of women who consumed 3 meals from a buffet under observation and completed a recall using ASA24-2016, those who were assisted by a trained paraprofessional ($n = 154$, 73.5% accuracy) did not have significantly more true matches to observed intake than participants who completed the recall independently ($n = 148$, 71.9% accuracy). Although women were able to report intake using ASA24 fairly accurately with low researcher burden, automated tools introduce the need for participant computer literacy (6). Another mixed-method study assessing the usability of ASA24 in a low-income population found a number of issues that affected participants success rates in independently completing a recall (7). The moderated group was more successful in completing the recalls compared with semimoderated and unmoderated groups. Usability issues were determined in the moderated and semimoderated groups and the issue with the highest proportional frequency was "Search Item Missing/Inaccurate." Inability to find an item through the search tool that matches what the participant is attempting to report might incline them to write in a text response. In our study, some corrections to text responses were made for items that exist in the FNDDS database, suggesting that participants may have been unable to find them in their search. However, we did not find a difference between the proportion of modifications made to supervised (assisted) versus home (unassisted) recalls. This suggests the provision of assistance did not significantly decrease participants' likelihood to write in a text response that could result in a suboptimal match.

When participants choose "Unfound foods" during the recall, they complete a series of questions to attempt to characterize what they are reporting. After the recall is completed and submitted, the best possible automated coding match is assigned. However, on the user interface side, participants are not informed of what ASA24 has determined to match before submission. For example, in one case, a participant attempted to report a branded "Perfect Bar with nuts and fruit," which was coded as pizza with meat. In follow-up prompts, the participant selected the "Unknown Food Kind" as "Mixtures" suggesting the automated coding matched it with a mixed dish. This was corrected to a granola bar during cleaning, but it would be preferable if participants had an opportunity to review the proposed "Unfound food" match and revise if needed prior to submitting. This may reduce the incidence of obviously unfit matches and preclude some of the need for posthoc data cleaning of open-ended text responses. For a food reported as an "Other, specify" text response, this may not be a useful solution as default foods are assigned based on the most common responses for the food reported in NHANES What We Eat in America (4). It may be advised for investigators to evaluate "Other, specify" responses, particularly for food groups of interest.

Part of the reason "beverage_milk" has the fourth highest number of modifications can be attributed to use of both 2014 and 2016 ASA24 versions (Figure 2). The 2007-2008 FNDDS database underlying ASA24-2014 did not include "almond milk," although participants wrote it in for text responses. The NCI provided SAS code to update 2014 food groups to the 2016 versions. To create a cohesive dataset, this was utilized and write-ins for "almond milk" from the 2014 version were corrected using 2011-2012 FNDDS (ASA24-2016) nutrient data. With releases of new ASA24-2018 and 2020 versions, updated databases and

user interfaces may decrease the need for text modifications necessitated by foods missing from older databases. However, this should be a consideration for investigators analyzing data collected using older versions of ASA24, particularly if dairy foods are of interest. Each version of ASA24 does not use the most recent FNDDS database available (ASA24-2018 uses FNDDS 2011–2012 supplemented with common foods in FNDDS 2013–2014, ASA24-2020 uses FNDDS 2015–2016 although FNDDS 2017–2018 is available) as FNDDS is updated every 2 years based on the most recent NHANES (15).

Past validations of dietary intake have used administration of doubly labeled water as a biomarker for energy intake. A study by Park et al. observed underreporting by 15–17% in participants that completed multiple ASA24 recalls (16). This trend of underreporting may explain low levels of correlation between reported mean energy intake and TEE in our study. The effect of data modifications was negligible, no significant differences in correlations were found before and after, although we observed a slightly higher correlation of modified data with energy intake from food and supplements (Table 7). Predictive energy requirements calculated by Harris-Benedict, Mifflin St. Jeor, and Cunningham calculations had a slightly higher correlation with modified mean energy intake compared with raw (Table 7A). This is not observed when considering only participants with at least 1 modified recall, as Harris-Benedict and Mifflin St. Jeor predictions correlate more highly with raw energy intakes (Table 7B). Based on the cross-sectional nature of the current study, without longitudinal measurements of TEE and reported energy intake, high correlation is not expected. TEE is an unbiased measure of energy intake for individuals in energy balance, maintaining their weight (17). Subjects were not asked whether they were trying to lose or gain weight. Thus, limitations of study design along with misreporting may have contributed to finding no meaningful difference between raw and modified data for correlation with TEE.

The Zimmerman article concluded that mean changes in nutrient intakes were not significant enough to warrant recommendations to clean data. However, differences were larger for individuals whose data were modified during cleaning. Relative interest in individual intake versus group intake can be a deciding factor in whether to make corrections to suboptimal matches. This prompts inquiry into best practices for dietary data collection for precision nutrition, tailored recommendations for individuals and population subgroups (18), with data cleaning as described here as a potentially advantageous strategy when using validated automated tools like ASA24. Discussions of automated dietary data collection tools have called for techniques to evaluate individual intake for investigating approaches to promoting optimal health. Improvement of individual nutritional status throughout the lifespan will require accurate individual nutritional assessment (19). Recalls improve the ability to collect quantitative individual intake data compared with FFQs, and ASA24 improves the efficacy of collecting and cleaning recall data compared with manual collection and coding (20). Although limited to a few nutrients and methodologies, our findings suggest that data cleaning modifications may be integral to these efforts, as they can change nutrient intakes, although not to a large scale in this cohort.

Limitations of this study include the number of recalls collected are not sufficient to estimate individual usual intake without further statistical modeling (21). Accordingly, correlational analysis of energy intake with TEE may not reflect accuracy of individual intake. Data col-

lection for this study occurred over several years, which necessitated the use of 2 versions of ASA24 and nutrient databases that were not up to date with recent food trends. In addition, our approach to modifying suboptimal coding of text responses was targeted for macronutrients and fiber and may have overlooked issues that affect micronutrients or food group intakes. Future studies could further explore the effect of data cleaning modifications with longitudinal dietary assessment design and statistical modeling to evaluate individual usual intake.

We found that modifications made to recalls during data cleaning changed mean nutrient intakes, in agreement with previous work by Zimmerman et al. We provided descriptions and categorization of food types that required modification from suboptimal automated coding, finding mismatches between animal and plant food products may explain changes to energy, protein, and fat intakes. As newer versions of ASA24 are released and search functionality improves, there may be less need to review and correct open-ended text responses. For investigators analyzing data collected using ASA24-2014 and 2016, it may be pertinent to consider modifying recalls for qualitative analysis based on foods and nutrients of interest, however, modification did not significantly impact individual energy intake correlation with TEE in this analysis.

Acknowledgments

We acknowledge Evelyn Holguin, Barbara Gale, Danna Juarez Rios, Justin Waller, Ashley Tovar, Christine Bowlus, Yanhua Li, Anna O'Dwyer, and Diane Han for human studies and physiology assessment in the USDA Nutritional Phenotyping Study and Zeynep Alkan for help with formatting supplementary data.

The authors' contributions were as follows—YYB, JEA, ELB, and CBS: designed research; YYB, JEA, EC, ELB, AK, and CBS: conducted research; AK: provided essential data material; YYB and DGL: analyzed data; YYB and JEA: wrote the manuscript; JEA, ELB, EC, AK, NLK, DGL, and CBS: revised the manuscript; CBS: had primary responsibility for final content; and all authors: read and approved the final manuscript.

References

1. Raper N, Perloff B, Ingwersen L, Steinfeldt L, Anand J. An overview of USDA's Dietary Intake Data System. *J Food Compos Anal* 2004;17(3):545–55.
2. National Cancer Institute Automated Self-Administered (ASA24) Dietary Assessment Tool. [Internet]. [Cited 2020 Dec 4]. Available from: <https://epi.grants.cancer.gov/asa24>.
3. Zimmerman TP, Hull SG, McNutt S, Mittl B, Islam N, Guenther PM, Thompson FE, Potischman N, Subar AF. Challenges in converting an interviewer-administered food probe database to self-administration in the National Cancer Institute Automated Self-Administered 24-Hour Recall (ASA24). *J Food Compos Anal* 2009;22(Supplement 1):S48–51.
4. Zimmerman TP, Potischman N, Douglass D, Dixit-Joshi S, Kirkpatrick SI, Subar AF, McNutt S, Coleman LA, Alexander GL, Kushi LH, et al. The effect of editing open-ended text responses on nutrient and food group estimates from the Automated Self-Administered 24-Hour Dietary Recall (ASA24). *Procedia Food Sci* 2015;4:160–72.
5. Kirkpatrick SI, Subar AF, Douglass D, Zimmerman TP, Thompson FE, Kahle LL, George SM, Dodd KW, Potischman N. Performance of the Automated Self-Administered 24-hour Recall relative to a measure of true intakes and to

- an interviewer-administered 24-h recall. *Am J Clin Nutr* 2014;100(1):233–40.
6. Kirkpatrick SI, Guenther PM, Douglass D, Zimmerman T, Kahle LL, Atoloye A, Marcinow M, Savoie-Roskos MR, Dodd KW, Durward C. The provision of assistance does not substantially impact the accuracy of 24-hour dietary recalls completed using the Automated Self-Administered 24-H Dietary Assessment Tool among women with low incomes. *J Nutr* 2019;149(1):114–22.
 7. Kupis J, Johnson S, Hallihan G, Olstad DL. Assessing the usability of the Automated Self-Administered Dietary Assessment Tool (ASA24) among low-income adults. *Nutrients* 2019;11(1):132.
 8. Thompson FE, Dixit-Joshi S, Potischman N, Dodd KW, Kirkpatrick SI, Kushi LH, Alexander GL, Coleman LA, Zimmerman TP, Sundaram ME, et al. Comparison of interviewer-administered and Automated Self-Administered 24-Hour Dietary Recalls in 3 diverse integrated health systems. *Am J Epidemiol* 2015;181(12):970–8.
 9. Subar AF, Kirkpatrick SI, Mittl B, Zimmerman TP, Thompson FE, Bingley C, Willis G, Islam NG, Baranowski T, McNutt S, et al. The Automated Self-Administered 24-hour dietary recall (ASA24): a resource for researchers, clinicians, and educators from the National Cancer Institute. *J Acad Nutr Diet* 2012;112(8):1134–7.
 10. Dao MC, Subar AF, Warthon-Medina M, Cade JE, Burrows T, Golley RK, Forouhi NG, Pearce M, Holmes BA. Dietary assessment toolkits: an overview. *Public Health Nutr* 2019;22(3):404–18.
 11. Tooze JA, Kipnis V, Buckman DW, Carroll RJ, Freedman LS, Guenther PM, Krebs-Smith SM, Subar AF, Dodd KW. A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Stat Med* 2010;29(27):2857–68.
 12. Baldiviez LM, Keim NL, Laugero KD, Hwang DH, Huang L, Woodhouse LR, Burnett DJ, Zerofsky MS, Bonnel EL, Allen LH, et al. Design and implementation of a cross-sectional nutritional phenotyping study in healthy US adults. *BMC Nutr* 2017;3(1):79.
 13. National Cancer Institute Reviewing and Cleaning ASA24 Data. [Internet]. [Cited 2020 Dec 4]. Available from: <https://epi.grants.cancer.gov/asa24/resources/cleaning.html>.
 14. ASA24 Known Issues and Workarounds. [Internet]. [Cited 2020 Dec 4]. Available from: <https://epi.grants.cancer.gov/asa24/resources/issues.html>.
 15. Food and Nutrient Database for Dietary Studies (FNDDS). [Internet]. [Cited 2020 Dec 10]. Available from: <http://www.ars.usda.gov/nea/bhnrc/fsrg>.
 16. Park Y, Dodd KW, Kipnis V, Thompson FE, Potischman N, Schoeller DA, Baer DJ, Midthune D, Troiano RP, Bowles H, et al. Comparison of self-reported dietary intakes from the Automated Self-Administered 24-h recall, 4-d food records, and food-frequency questionnaires against recovery biomarkers. *Am J Clin Nutr* 2018;107(1):80–93.
 17. Subar AF, Kipnis V, Troiano RP, Midthune D, Schoeller DA, Bingham S, Sharbaugh CO, Trabulsi J, Runswick S, Ballard-Barbash R, et al. Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the OPEN Study. *Am J Epidemiol* 2003;158(1):1–13.
 18. de Toro-Martín J, Arsenault BJ, Després J-P, Vohl M-C. Precision nutrition: a review of personalized nutritional approaches for the prevention and management of metabolic syndrome. *Nutrients* 2017;9(8):913.
 19. Péter S, Saris WHM, Mathers JC, Feskens E, Schols A, Navis G, Kuipers F, Weber P, Eggersdorfer M. Nutrient status assessment in individuals and populations for healthy aging – statement from an expert workshop. *Nutrients* 2015;7(12):10491–500.
 20. Amoutzopoulos B, Steer T, Roberts C, Cade JE, Boushey CJ, Collins CE, Trolle E, de Boer EJ, Ziauddeen N, van Rossum C. Traditional methods v. new technologies – dilemmas for dietary assessment in large-scale nutrition surveys and studies: a report following an international panel discussion at the 9th International Conference on Diet and Activity Methods (ICDAM9), Brisbane. *J Nutr Sci* 2018;7:e11.
 21. Conrad J, Nöthlings U. Innovative approaches to estimate individual usual dietary intake in large-scale epidemiological studies. *Proc Nutr Soc* 2017;76(3):213–9.