Reviewing and Cleaning ASA24® Data

ASA24 data may be reviewed to determine if the analysis files contain missing data, incorrect matches for text entries, or outliers. Any of these may require recoding and reanalysis of the data, which must be done outside the ASA24 system. The following information presents recommended procedures for reviewing and cleaning ASA24 data. The time and effort that researchers choose to allocate to reviewing and cleaning data may depend on the size of the study and the research questions.

The following ASA24 output files will be referred to by their acronyms:

Version	Analysis File	Content Description
ASA24-2016	Responses	Food and supplement names from the Quick List, probe questions and answers.
	Items	Food and Nutrient Database for Dietary Studies (FNDDS) Food Codes, Gram weights, Food Pattern Equivalents (FPED) for each item reported.
	Totals	Daily Total of FNDDS nutrients and FPED food groups for a consumption day
	INS	Individual Supplements Analysis File – Supplement Codes with their nutrients for each supplement reported – includes only those nutrients found in FNDDS
	TS	Daily Total Supplements Analysis File – total nutrient intake from all supplements reported on a consumption day – includes only those nutrients in FNDDS
	TNS	Daily Total Nutrients from Foods and Supplements Analysis File – FNDDS nutrients from all foods and supplements reported on a consumption day

Version	Analysis File	Content Description
ASA24-2014, ASA24-Kids-2104, ASA24-Canada-2014 ASA24-2011, ASA24-Kids-2012	MS	MySelections Analysis File - Food and supplement names from the Quick List, probe questions and answers.
	INF	Individual Foods Analysis File – Food and Nutrient Database for Dietary Studies (FNDDS) (Canadian version uses Canadian Nutrient File (CNF)) food codes, gram weights and nutrients for each food or beverage reported
	TN	Daily Total Nutrients Analysis File - FNDDS (CNF for Canada) nutrients from all foods in a given day for a consumption day
	INFMYPHEI	Individual Foods MyPyramid (MPED) Healthy Eating Index (HEI) Analysis File – FNDDS Food Codes, Gram weights,

		MPED and HEI Whole Fruit variable for each food or beverage reported (Note: MPED not available for Canadian version.)
	TNMYPHEI	Daily Total Nutrients and MPED Analysis File – FNDDS (CNF for Canada) nutrient and MPED food group and HEI Whole Fruit variable for each consumption day. (Note: MPED not available for Canadian version.)
	INS	Individual Supplements Analysis File – Supplement Codes with their nutrients for each supplement reported – includes only those nutrients found in FNDDS
	TS	Daily Total Supplements Analysis File – total nutrients from all supplements reported on a consumption day – includes only those nutrients found in FNDDS
	TNS	Daily Total Nutrients from Foods and Supplements Analysis File – FNDDS (CNF for Canada) nutrients from all foods, beverages and supplements reported on a consumption day

1. Missing data: In the ASA24 system, rows in the *Items/INF/INFMYPHEI/INS* files with no kcal, nutrients, or other components should be examined to determine the extent of missing data. As a guideline, the National Center for Health Statistics (NCHS) guidelines for using the National Health and Nutrition Examination Survey (NHANES) data state that it is generally acceptable to use data if 10% or less of the data for a variable is missing (http://www.cdc.gov/nchs/tutorials/NHANES/Preparing/CleanRecode/Info1.htm).

   Missing data may be due to one of the following:

   a. Errors in the ASA24 database: there are known errors in the ASA24 database (particularly in the Beta version) that result in missing data. Summaries of these errors and their workarounds may be found in the Known Issues & Workarounds information on http://epi.grants.cancer.gov/asa24/resources/issues.html .

   b. Breakoffs: if a respondent enters foods during the Quick List (reflected in the MS file) but leaves the ASA24 recall prior to reaching the final question, some or all of the rows in the *Items/INF/INFMYPHEI/INS* files may be missing data. Researchers will need to decide whether or not to include recalls that are breakoffs (i.e., recall started but not completed so that details are missing for some or all foods and drinks reported); if included, each row needs to be coded (by applying a default food code and portion code) and analyzed outside the ASA24 system.

   c. No ingredient: Many foods in the ASA24 system are represented by multiple rows in the *Items/INF/INFMYPHEI* files. For example, a turkey sandwich will have one row for the bread, one row for the turkey, one row for mayonnaise, and so on. In many cases, when a respondent reports that they did not know if an item was on their sandwich (e.g., they didn't know if there was cheese on their sandwich), no food code is applied and there is an

empty row in the INF/INFMYPHEI files. There is no recoding needed for these – they may be ignored.

2. Text entries: Within the ASA24 software, respondents are given two opportunities to enter open-ended text: "other" and "match not found."

    a. "Other" is available as a response to questions about food details, such as brand name or cooking method. The system collects but does not use the text response, instead assigning a default food code from the Food and Nutrient Database for Dietary Surveys (FNDDS) based on intake data from NHANES. For example, when reporting a green salad, a respondent may select "Other" as the response to the kind of vegetables in the salad and enter turnip in the text field; the system assigns the default code selected for the ASA24 system, which for vegetables on a salad, is tomato.

    b. "Match not found" can be selected by respondents if they cannot find a food or drink they want to report. After the respondent enters a text response to describe the food or drink, the ASA24 system asks a series of general questions to better identify the item, including the food category. Based on this information, a default food code from FNDDS is assigned (i.e., a food code used when details are not known, such as "Bread, not specified as to commercial or homemade"); when a default FNDDS food code is not available for that food, a food code is selected based on intake data from NHANES. For example, if a respondent enters "oatmeal brownies," the ASA24 system asks "what kind of food was it?" then, if the respondent selects "breads, other baked goods," the ASA24 system asks "what kind was it?" If the respondent then selects "brownies," a food code for chocolate brownies is assigned to the item.

    In large studies, the review and correction of codes applied to "Other" and "Match not found" entries would be time and resource consuming and possibly not feasible. Preliminary analysis of the impact of review and correction of free text in a study of 1,200 participants suggests that this level of cleaning and recoding may not be necessary (Zimmerman TP et al., *Procedia Food Science* 2015 (4): 160 – 172). However, researchers should consider whether or not to review their data at this level based on the research question and level of precision required of the data.

    In smaller studies and clinical settings, Researchers may wish to review and correct, if necessary, food codes applied to "Other" and "Match not found" entries to ensure consistency with what the respondent appeared to be reporting. Whether or not to complete this step may depend on the research question and the precision required of the data.

3. Outlier review: Outlier reports using established cut off points (Appendix A) for portion size or energy/nutrient quantity can identify intakes with unusually high or low portions or nutrient amounts. Portion and nutrient outlier reports are helpful in finding errors either in the ASA24 database or possible respondent errors. This review is somewhat subjective, but some obvious errors may be found. However, caution should be exercised in discarding recalls with high or low intakes since intakes of energy, nutrients and food groups fluctuate from day to day.

4. Example of potential error uncovered by examining portion and nutrient outliers: when respondents report items like sandwiches and tacos that they prepare themselves, they are asked amounts of each ingredient in the sandwich/taco, and then asked how many of the sandwiches/tacos they ate. When reporting the amount of each ingredient in each sandwich/taco, some respondents who ate more than one of the sandwich/taco have reported what appears to be the total amount of the ingredient rather than the amount per sandwich/taco. For example, a respondent reports two tacos, and reports 1 cup of ground beef when asked how much ground beef per taco. This results in 2 cups of ground beef being entered; in this case, the unrealistic amount of ground beef per taco leads to the suspicion that the respondent actually reported the total amount of ground beef eaten, rather than the ½ cup of ground beef on each of the tacos eaten.
5. In addition to the above outlier review, statistical thresholds can be used to identify and examine outliers. For example, intakes above the 75th percentile plus two or three times the interquartile range might be flagged for review. Researchers should carefully review recalls for outliers and exercise caution in discarding data, as noted above, since it is possible to have a high or low intake of energy, nutrients, or food groups on a given day. For example, review of recall data for a given respondent with a high intake of Vitamin A might reveal consumption of carrots or other foods high in vitamin A on the recall day.
6. Duplicate entries: Respondents sometimes enter what appear to be duplicate entries for a single food, possibly reported in two different ways. For example, a respondent may add the various components of a turkey sandwich and then also add the food "turkey sandwich." This can result in two turkey sandwiches in the QuickList (i.e., MS file) if not corrected by the respondent. Finding this type of error requires a visual review of the MS file to determine if foods reported within a single meal appear to be duplicates. Again, this is a subjective review and caution is warranted in modifying data. In smaller studies or clinical settings, it may be possible to follow up with the Respondent to verify what was actually consumed.

Appendix A
Criteria Used to Triage the Records Requiring a Review for Accuracy

Portion outliers: The following quantities exceed usual portions for items consumed at one eating occasion.

- Beverages greater or equal to ½ gallon
- Meat, fish, poultry greater than or equal to 12 ounces (342 grams)
- Mixed dishes greater than or equal to 6 cups
- Snack foods (chips, nut, etc.) greater than or equal to 8 ounces by weight

Nutrient outliers: Cut points are based on the 5th and 95th percentile of intakes from NHANES data.

1. Kcal

Gender/Age	Low	High
Adult women >/=12 years old	600	4400
Adult males >/= 12 years old	650	5700

2. Protein

Gender/Age	Low	High
Adult women >/=12 years old	10	180
Adult males >/= 12 years old	25	240

3. Fat

Gender/Age	Low	High
Adult women >/=12 years old	15	185
Adult males >/= 12 years old	25	230

4. Vitamin C

Gender/Age	Low	High
Adult women >/=12 years old	5	350
Adult males >/= 12 years old	5	400

5. Beta-carotene

Gender/Age	Low	High
Adult women >/=12 years old	15	7100
Adult males >/= 12 years old	15	8200