HCHS/SOL Dietary Data Overview, Methods and Guidelines

June 2013
Version 1.1

Prepared by
HCHS/SOL Coordinating Center
Collaborative Studies Coordinating Center
UNC Department of Biostatistics

Daniela Sotres-Alvarez
Anna Maria Siega-Riz
Nathan Gotman
Natalia Gouskova

Please send questions, suggestions and comments to dsotres@unc.edu

Table of Contents

1 Updates in INV4.2 DIET data release

Some diet datasets and two diet scores are being released for the 1st time to HCHS/SOL investigators in June 2013 (INV4.2). As a consequence the Dietary Data Overview, Methods and Guidelines Document is updated.

MAIN updates in Version 1.1 (June 2013)

- **Seven NDSR files are distributed for the 1st time:** 2, 7, 14, 16, 17, 18 and 19 (Table 4.1).
- **Dataset FOOD_GROUP_DERV** includes servings/day of ad hoc food groups (My Pyramid and other *ad hoc* food groups; Appendix A) at the recall level.
- **Dataset PRED_NUTR_DERV** includes predicted nutrient usual intake by NCI method
- **Two overall dietary scores DIET_Score_JAMA and The Alternative Healthy Eating Index (AHEI2010)** were added to the Participant Derived Dataset (**PART_DERV)**. Documentation for these variables is in Section 18 "DIET" of the Derived Variable Dictionary.
- Appendices B and C include SAS code to clean dietary recalls and to predict energy usual intake by NCI method using HCHS/SOL data.

MAIN updates in Version 1.0 (Dec 2012)

- **Four NDSR files are distributed for the 1st time**: 3, 8, 13 and 15 (Table 4.1).
- **Three NDSR files that have been previously released have been updated**: DTIA (NDSR4), DTSA (NDSR9) and S24A (NDSR12). One important change is that key identifiers' names for DTSA, DTIA and S24A have changed (see section 4.1 for details).

2 FORWARD

Note to User of these HCHS/SOL Dietary Data Overview, Methods and Guidelines:

This document is not intended for direct citation or distribution outside of the immediate HCHS/SOL Study. It should be considered confidential and proprietary to HCHS/SOL investigators.

- **Diet Data Overview Section 3** presents a summary of dietary intake assessment conducted during HCHS/SOL baseline exam and one-year follow-up call. **Section 4** describes SAS datasets from 24-hr dietary and supplement recalls and 30-day supplement recall.
- **Methods Sections 5 to 7** describe HCHS/SOL dietary data (24-hour recalls and Food Propensity Questionnaire) applying National Cancer Institute (NCI) method and NCI macros to estimate or predict usual intake (nutrients or foods).
- **Sections 8 and 9** present guidelines, resources and recommendations for use of the data in analysis and presentation for publications.
- In **appendix C** we include SAS code to execute the NCI macros using HCHS/SOL data

3 Dietary Intake Assessment in HCHS/SOL at Baseline

As part of the HCHS/SOL baseline exam, dietary intake was assessed with two 24-hour dietary recalls. Manuscript #8 “Dietary assessment methodology in diverse Latino groups: Hispanic Community Health Study/Study of Latinos” will document descriptively methodological concerns and how they were addressed for HCHS/SOL. At one-year follow-up call the Food Propensity Questionnaire (FPQ) was administered. Manuscript #39 “Reduction of elements in a food frequency questionnaire for estimation of usual food intake: An example from the Hispanic Community Health Study” will document how the FPQ was adapted for HCHS/SOL for the purpose of implementing the NCI methodology.

3.1 Two 24-hour dietary and supplements recalls (24-hr and 30-day)

- **First dietary recall** was administered in-person at the field center.
- **Second dietary recall** was mostly telephone administered, and conducted at least five days and ideally within 45 days, following the initial examination interview.
- **Both 24hr dietary recalls were conducted using the NDSR software** developed by the Nutrition Coordinating Center (NCC) at the University of Minnesota which employs the multiple pass method. For the first dietary recall food models were used, and an amounts booklet was provided for estimating portion sizes in the 2nd recall. NDSR version 2011 (http://www.ncc.umn.edu/index.html) contains over 18,000 foods, 8,000 brand name products, many ethnic foods supplements and vitamins. The program provides values for 158 nutrients, nutrient ratios, and other food components. NDSR calculates nutrient intake and presents the data in several formats, including daily nutrient totals, nutrient amounts per individual food, daily totals compared to the Dietary Recommended Intakes (DRIs) and a new food group serving count system.
- **The NDSR Dietary Supplement Assessment Module (DSAM)** was used for the 24-hour dietary recall. This assessment utilizes the most currently available NHANES Supplement Database with enhancements from NCC and allows for the collection of 24-hour and/or 30-day intake of all dietary supplements and antacids. The goal of the dietary supplement recall is to assess use of all types of dietary supplements and over-the-counter antacids. Over-the-counter antacids are included in this assessment because many of these products contain calcium. **For the in-person interview**, the period covered for dietary supplement intake is the same time period covered by the 24-hour dietary recall and the past 30 days. **For the telephone (or second interview)** the DSAM was limited to the 24-hour dietary recall, and not past 30 day supplement use.
- For more detail on 24-hour dietary recall, 24-hour and 30 day supplement use assessment procedures, see HCHS/SOL Manual of Operations 11 “Diet and Supplements”.

3.2 Food Propensity Questionnaire (FPQ)

- The FPQ assesses intake of specific foods and food groups during the past 12 months. It was administered at annual follow up year 1.
- On April 2010 the FPQ was shortened in order to reduce participant burden (time)
    - FPQ long (version A): 139 food items from 228 individual questions.
    - FPQ short (version B): 115 food items from 137 individual questions.
- Manuscript #39 “Reduction of elements in a food frequency questionnaire for estimation of usual food intake: An example from the HCHS/SOL” by Catellier et al is a methodological manuscript that will present a strategy and rationale for dropping a subset of questions from the National Cancer Institute (NCI) Food Propensity Questionnaire (FPQ) such that the

instrument maintains its utility in estimating usual food intake in conjunction with repeated 24h dietary recalls without loss of accuracy.

3.3 Dietary Behavior Questionnaire (DBEA)

Two questions were included on the Dietary Behavior Questionnaire (DBEA) that solicited information regarding the overall types of foods being consumed belonging more to traditional diets or that of the US population and the second on the frequency and location of foods consumed away from home.

4 SAS Datasets from 24-hr Dietary Recalls and Supplement Recalls (24-hr and 30-day)

Data release INV4 includes raw dietary data from 24-hr dietary recalls, 24 hour and 30 day supplement recalls, derived food groups, a derived dataset with predicted usual intake for 20 nutrients, and the Alternative Healthy Eating Index 2010. Data from the Food Propensity Questionnaire (FPQ) is included in the annual follow-up data release (AFUINV1).

4.1 24-hr Dietary and Supplement Recalls and 30-day Supplement Recalls (NDSR files)

NDSR files 2-4, 7-9, and 12-19 are raw data files for the 24 hour dietary and supplement recalls and the 30-day supplement recall. These are multiple-record-per-participant datasets (Table 4.1.1). NDSR files 5, 6, 10, 11, 20 and 21 were not used in HCHS/SOL. The included files have been reformatted from their original NDSR format to be consistent with HCHS/SOL dataset and variable naming conventions (see HCHS/SOL Investigator Use Database Overview). Each dataset was given a three letter abbreviation and a version index (with A being the first version of a form). **The variables were named using the form name and a variable index, where the variable index corresponds to the column number in the NDSR manual.** The SAS files contain variable labels with English descriptions of each variable.

The NDSR version used to collect the data for a participant can be found in DTIA17 (ranges from 2007-2010). All raw files were processed using version 11 (2011) of the NDSR software, which uses 2011 USDA database of nutrient contents of foods. **For information on NDSR Food Group Serving Count System and serving sizes see Appendix 10 of the NDSR 2011 Manual.**

Among the 16,415 HCHS/SOL participants, 99% (16,285) have the 1st dietary recall (in-person at clinic visit) and 94% (15,424) have the 2nd recall (most conducted by telephone) (Table 4.1.2). The percentage of recalls referring to a weekend day (Friday to Sunday) varies among sites for 1st dietary recalls, 58% for Chicago and 23% for Miami. In contrast, the percentage of 2nd recalls referring to a weekend day is around 25% for all sites other than the Bronx which had 34% of 2nd recalls refer to a weekend day. For the first recall, 71% self-reported that dietary intake was their usual amount, 9% a lot more than usual, and 20% a lot less than usual.

For assessment of dietary intake with multiple 24-hr dietary recalls, statistical models provide a better estimator of usual intake than simply averaging 24-hr dietary recalls. See section 5 for one statistical method to summarize both 24-hr recalls.

Table 4.1.1. SAS datasets from dietary and supplement recalls (NDSR raw data)

NDSR File	HCHS/SOL dataset	Dataset description	Key field 1	Key field 2[1]	Key field 3[1]
2	DIEA	Nutrients at the whole food level	ID	RECALLNUM	FOODID
3	MEOA	Nutrients at the meal level	ID	RECALLNUM	MEALID
4	DTIA	Nutrients at the daily totals level	ID	RECALLNUM	
7	FSCA	Food groups at the whole food level	ID	RECALLNUM	FOODID
8	MSCA	Food groups at the meal level	ID	RECALLNUM	MEALID
9	DTSA	Food groups at the daily totals level	ID	RECALLNUM	
12	S24A	Total 24 hour supplement intake	ID	RECALLNUM	
13[2]	SMIA	Averaged 30-day supplement intake	ID		
14	P24A	Product file for 24 hour supplement intake	ID	RECALLNUM	PRDID
15[2]	PMIA	Product file for 30-day supplement intake	ID	SUPPLID	
16	I24A	Product ingredients for 24 hour supplement intake	ID	RECALLNUM	INGID
17[2]	I30A	Product ingredients for 30 day supplement intake	ID	INGID	
18	B24A	Blend ingredients for 24 hour supplement intake	ID	RECALLNUM	BLDID
19[2]	B30A	Blend ingredients for 30 day supplement intake	ID	BLDID	

NDSR files 5, 6, 10, 11, 20 and 21 were not used in HCHS/SOL
[1]Descriptions of key field variables are as follows:
RECALLNUM – first or second 24 hour dietary recall
FOODID – Food file ID. MEALID plus a 3 digit index (ABBB or AABBB)
MEALID – 1-2 digit index for meal or eating occasion
PRDID or SUPPLID – 1-3 digit index for DSAM product file ID.
INGID – Ingredient ID. PRDID plus a 3 digit index (ABBB, AABBB, or AAABBB)
BLDID – Blend ingredient ID. INGID plus a 3 digit index (AAAABBB, AAAAABBB, or AAAAAABBB)
[2]30 day supplement intake was assessed only at the clinic visit. RECALLNUM was not needed as a key field.

4.2 Food Groups (FOOD_GROUP_DERV)

The SAS multiple-record-per-participant dataset FOOD_GROUP_DERV contains derived variables for 50 broad food groups at the daily level created from DTSA (165 NDSR food codes at the daily level). Description of these 50 derived variables can be found in appendix A. Serving counts for individual NDSR food codes were added to create serving counts per day for each food group. The dataset contains up to 2 observations per participant, depending on how many 24hr dietary recalls were collected, reliable according to interviewer (DTIA16) and cleaned at the 24hr recall level based on daily energy intake (DTIA20). See appendix B for details on data cleaning processes.

For assessment of dietary intake with multiple 24-hr dietary recalls, statistical models provide a better estimator of usual intake than simply averaging 24-hr dietary recalls. See section 5 for the NCI method to summarize both 24-hr recalls allowing incorporating the FPQ which in particular can improve estimates for episodically consumed foods or food-groups. The FPQ assesses long-term dietary intake and captures 'true' non-consumers. In contrast, 24-hour recalls are short-term instruments and overestimate the percent of non-consumers because of the large day-to-day variability within persons.

Table 4.1.2 Characteristics of 24-hour recalls by sequence and site

		1st recall (In-person)					*2nd recall (Telephone)*				
Variable	*Category*	*Bronx*	*Chicago*	*Miami*	*San Diego*	*Overall*	*Bronx*	*Chicago*	*Miami*	*San Diego*	*Overall*
HCHS cohort with recall data	N	4,062	4,122	4,058	4,043	16,285	3,665	3,989	3,872	3,898	15,424
Intake Reliability	Reliable	3,989	4,071	3,923	3,957	15,940	3,597	3,925	3,760	3,726	15,008
	Unreliable - cannot recall meal(s)	35	11	36	29	111	20	10	16	36	82
	Unreliable - other reasons	38	40	99	57	234	48	54	96	136	334
% weekend (Fri, Sat, Sun)	(%)	25.4	58.2	23.1	33.1	35.0	33.5	23.2	20.7	25.6	25.6
Interview administrated	Missing	5.9	4.1	6.3	7.0	5.8	2.2	3.5	3.9	4.3	3.5
	In-person	94.1	95.9	93.5	92.9	94.1	14.3	3.0	10.1	32.2	14.9
	Telephone		0.05 (2)	0.17 (7)	0.12 (5)	0.09 (14)	83.6	93.5	86.0	63.5	81.7
Interview type	Missing	6.0	4.1	6.3	6.8	5.8	2.2	3.5	3.9	4.2	3.5
	Scheduled	93.9	95.8	93.3	92.7	93.9	20.3	3.6	9.0	12.7	11.2
	Unscheduled	0.15 (6)	0.10 (4)	0.39 (16)	0.57 (23)	0.30 (49)	77.5	92.9	87.1	83.1	85.3
Self-report usual amount	Usual	59.7	77.6	72.9	74.0	71.1	69.2	83.8	78.5	78.2	77.6
	Considerably more than usual	10.8	7.1	6.1	10.0	8.5	8.2	4.7	4.0	8.5	6.3
	Considerably less than usual	29.5	15.4	21.0	16.0	20.4	22.5	11.4	17.5	13.3	16.1

4.3 Predicted Nutrient Usual Intake (PRED_NUTR_DERV)

The predicted nutrient usual intake file (PRED_NUTR_DERV) contains derived variables for predicted usual intake for 20 nutrients (out of 158 available from the 2011 NDSR; see Table 4.3). Nutrient intake was predicted from an amount model (i.e. a one-part nonlinear mixed model) specified by the NCI method (Tooze et al, 2010), using single component SAS macros developed at NCI http://riskfactor.cancer.gov/diet/usualintakes/macros.html. This method estimates the within and between person components and corrects for the high intra-individual variation intrinsic to 24-hr recalls given that individuals do not eat the same foods every day. We excluded recalls with daily energy intake (DTIA20) below the recall-gender specific 1st percentile or above the 99th percentile, and recalls that were unreliable according to the interviewer (DTIA16). See appendix B for details in data cleaning. **Participants with at least one clean dietary recall have predicted nutrient intake (n=16,172).** Models were adjusted for gender, age, Hispanic/Latino background, field center, weekend (including Friday), self-report intake amount (more, same or less than usual amount), and sequence (1st recall-conducted in person or 2nd recall majority conducted by phone). For a detailed description of this method and its implementation in HCHS/SOL see sections 5 and 6.

Table 4.3. Predicted nutrient intake variables available in PRED_NUTR_DERV

Variable Name	Label
PRED_ENERGY	Energy (kcal) - NCI predicted intake (based on DTIA20)
PRED_TOTFAT	Total Fat (g) - NCI predicted intake (based on DTIA21)
PRED_TOTCARB	Total Carbohydrate (g) - NCI predicted intake (based on DTIA22)
PRED_TOTPROT	Total Protein (g) - NCI predicted intake (based on DTIA23)
PRED_TOTSFA	Total Saturated Fatty Acids (SFA) (g) - NCI predicted intake (based on DTIA28)
PRED_TOTMUFA	Total Monounsaturated Fatty Acids (MUFA) (g) - NCI predicted intake (based on DTIA29)
PRED_TOTPUFA	Total Polyunsaturated Fatty Acids (PUFA) (g) - NCI predicted intake (based on DTIA30)
PRED_FIBER	Total dietary fiber (g) - NCI predicted intake (based on DTIA38)
PRED_VITC	Vitamin C (ascorbic acid) (mg) - NCI predicted intake (based on DTIA52)
PRED_FOLATE	Total Folate (mcg) - NCI predicted intake (based on DTIA58)
PRED_CA	Calcium (mg) - NCI predicted intake (based on DTIA60)
PRED_FE	Iron (mg)- NCI predicted intake (based on DTIA63)
PRED_NA	Sodium (mg) - NCI predicted intake (based on DTIA67)
PRED_K	Potassium (mg) - NCI predicted intake (based on DTIA68)
PRED_PCTFAT	% Calories from Fat - NCI predicted intake (based on DTIA119)
PRED_PCTCARB	% Calories from Carbohydrate - NCI predicted intake (based on DTIA120)
PRED_PCTPROT	% Calories from Protein - NCI predicted intake (based on DTIA121)
PRED_PCTSFA	% Calories from SFA - NCI predicted intake (based on DTIA123)
PRED_TOTTRANSFAT	Total Trans-Fatty Acids (TRANS) (g) - NCI predicted intake (based on DTIA132)
PRED_ VITA_RAE	Total Vitamin A Activity (Retinol Activity Equivalents) (mcg) - NCI predicted intake (based on DTIA165)

4.4 Diet scores (PART_DERV)

Two diet scores are included in the Participant Derived Dataset (PART_DERV) and documented in Section 18 of the HCHS/SOL Baseline Examination Derived Variable Dictionary. Diet scores were created using only 24hr recall data, and included only those that were reliable according to the interviewer (DTIA16) and clean at the 24hr recall level based on daily energy intake (DTIA20). See appendix B for details on data cleaning.

4.4.1 Diet_Score_JAMA

This score was calculated by assigning participants a score of 1 to 5 (with 5 being the most favorable quintile) according to sex-specific quintiles of predicted daily usual intake of saturated fatty acids, potassium, calcium and fiber. The four scores were summed and the highest 40 percentile (diet_score_JAMA_c2) was considered a healthier diet (Liu K, et al Circulation, 201; 125 (8): 996-1004). See HCHS/SOL Baseline Examination Derived Variable Dictionary for more information about Diet_Score_JAMA.

4.4.2 Alternative Healthy Eating Index (AHE2010)

The alternative healthy eating index (AHEI-2010) is a measure of diet quality based on foods and nutrients predictive of chronic disease risk (Chiuve SE et al, 2012). This section describes briefly how the AHEI-2010 was calculated using the 24hr dietary recalls in HCHS/SOL baseline (2008-2011); additional detail is provided in Section 18.3 of the of the HCHS/SOL Baseline Examination Derived Variable Dictionary.

AHEI-2010 score is derived using the following 11 components:
1) Vegetables without potatoes, servings/day
2) Whole Fruit (i.e. does not include fruit juice), servings/day
3) Whole grains, servings/day
4) Sugar sweetened beverages and fruit juice, servings/day
5) Nuts and legumes, servings/day
6) Red/processed meat, servings/day
7) Trans Fat, % energy
8) Long-chain (n-3) fats (EPA+DHA), mg/day
9) Polyunsaturated fatty acids (PUFA), % energy;
10) Sodium, mg/day;
11) Alcohol, drinks/day

AHEI-2010 score is the sum of the 11 individual components' scores, each with a range from 0 (worst) and 10 (best). Hence, **AHEI-2010 takes values from 0 to 110.** Higher scores represent healthy eating habits and lower scores represent unhealthy eating habits. Only scores for components "whole grains", sodium and "alcohol" are gender-specific.

In HCHS/SOL, AHEI-2010 was calculated from available 24hr dietary recall data (one or two reliable recalls per participant) using the NCI method to predict usual intake for each component. A participant has an AHEI-2010 score if he/she has at least one 24-hr dietary recall after exclusions. We excluded recalls with daily energy intake (DTIA20) below the sequence-gender specific 1st percentile or above the 99th percentile and recalls that were unreliable according to the interviewer (DTIA16, see appendix B for details).

We followed 4 general steps to calculate the AHEI-2010:
1. Quantify each of the 11 components at the 24 hour dietary recall level.
2. Combine 24 hour dietary recalls to predict usual intake of the component.
3. Score each of the 11 individual components according to the AHEI-2010 algorithm
4. Compute the AHEI-2010 score as the sum of scores for individual components.

1. Quantify each of the 11 components at the dietary recall level.

Table 1 in Section 18 of the HCHS/SOL Baseline Examination Derived Variable Dictionary has the specific NDSR food subgroups used to define each component of AHEI-2010. Most components are the sum of the NDSR food subgroups. When serving sizes for NDSR were different than those specified in Chuive et al, servings were rescaled. For example, the serving size in NDSR for citrus juice (DTSA4) is 4 fluid ounces whereas in Chiuve el al is 8 fluid ounces.

2. Combine both dietary recalls to predict usual intake of the component.

In HCHS/SOL, AHEI-2010 was calculated from available 24hr dietary recall data with one or two recalls per participant using the NCI method to predict usual intake for each component (see section 6 for model specification). Food groups (i.e. components 1 to 6 and 11) and long-chain (n-3) fats (EPA+DHA) were modeled with a two-part model with correlated random effects. Percent of energy from trans-fat, % energy from PUFA, and sodium were modeled with a one-part model (amount model). Predicted usual intake is calculated for all participants with at least one 24-hr recall after exclusions. We excluded recalls with daily energy intake (DTIA20) below the sequence-gender specific 1st percentile or above the 99th percentile and recalls that were unreliable according to the interviewer (DTIA16). See appendix B for details. Note that it is possible that a participant has only the second dietary recall included to obtain predicted usual intake if the 1st dietary recall is unreliable or extreme according to observed daily energy intake (DTIA20).

3. Score each component

The scores for each individual component were computed according to the formulas given in Table 2 in Section 18 of the HCHS Baseline Examination Derived Variable Dictionary, which are based on Table 1 in Chiuve et al. Intermediate intakes were scored proportionately between 0 and 10 (McCullough ML et al, 2002). The variable in column "**Score calculation"** is the predicted usual intake of that component using NCI method on available dietary recalls. For example, variable "PRED_VEG_WOPOT" is the predicted usual intake (servings/day) of component "Vegetables without potatoes". For sodium, the lowest and highest deciles of sodium intake were estimated using NCI macro DISTRIB. For whole grain, the criterion for maximum score was converted to servings/day units (5 and 6 servings/day for women and men, respectively) from grams/day units (as defined in Chiuve et al, 2012) to match the NDSR unit.

References

S E Chiuve et al. (2012) "Alternative Dietary Indices Both Strongly Predict Risk of Chronic Disease"; J. Nutr. 142: 1009 – 1018

M L McCullough et al. (2002) "Diet quality and major chronic disease risk in men and women: moving toward improved dietary guidance"; Am J. Clin. Nutr. 76: 1261 - 71

5 NCI Method to Estimate or Predict Usual Dietary Intake

National nutritional surveys estimate usual nutrient or food intake to assess dietary deficiencies and excesses and adherence to dietary recommendations. For assessment of dietary intake with multiple 24-hr dietary recalls, statistical models provide a better estimator of usual intake than simply averaging 24-hr dietary recalls. Dodd et al (Am Diet Assoc, 2006) provide an excellent review on several statistical methods for estimating usual intake: NRC (National Research Council, ISU (Iowa State University) (Nusser et al, 1996a), Best Power Method (Nusser et al, 1996b), and the NCI Method (Tooze et al, 2006). More recently, Souverin et al (Aur J Clin Nutr (2011) compared these methods with MSM (Multiple Source Method) (Haubrock et al, 2011) and SPADE (Statistical Program for age-adjusted dietary assessment) (Waijers et al, 2006; Deckers).

In this section, we describe the National Cancer Institute (NCI) method to estimate usual dietary intakes of foods and nutrients (http://riskfactor.cancer.gov/diet/usualintakes/method.html) which was used to predict usual nutrient intake in HCHS/SOL. This method can be used to:

- estimate the distribution of usual intake for a population or subpopulation;
- assess the effects of individual covariates on consumption; and
- predict individual usual intake for use in a model to assess the relationship between diet and disease or other variable.

The premise of the NCI method is that usual intake is equal to the probability of consumption on a given day times the average amount consumed on a "consumption day". The methods used for dietary components that are consumed nearly every day by nearly everyone differ slightly from those used for dietary components that are episodically consumed. In general, the former category (ubiquitously consumed or consumed daily by almost everyone) includes most nutrients whereas the latter category (episodically consumed) includes most foods, though there are exceptions. An excellent resource for learning about this methodology is the “Measurement Error Webinar Series” (http://riskfactor.cancer.gov/measurementerror/; see section 9 for more information).

5.1 MODEL

NOTATION

i	Index subject
j = 1,2	Index 24 hour dietary recall (24hr)
T_{ij}	True (unobserved) **nutrient or food-item** intake for individual i on day j (Original Scale)
R_{ij}	Dietary intake from 24hr
R_{ij}^{*}	Box-Cox transformation (with parameter λ) $R_{ij}^{*} = g(R_{ij}, \lambda) = \begin{cases} (R_{ij}^{\lambda} - 1)\lambda^{-1} & if\ \lambda \neq 0 \\ \ln(R_{ij}) & if\ \lambda = 0 \end{cases}$

Part I (Probability of consumption): $p_i = \Pr[T_{ij} > 0]$

Part II (Amount on a consumption day): $A_i = E(T_{ij} | T_{ij} > 0)$

Premise of NCI method: Usual Intake = Probability x Amount = p_i x A_i

Part I

$$logit\{p_i\} = \mathbf{x}_{\mathrm{Ii}}^{\mathrm{T}}\boldsymbol{\beta}_I + u_{Ii} \qquad u_{Ii} \sim N(0, \sigma_{u_I}^2)$$

Part II

$$g(R_{ij}, \lambda) = \mathbf{x}_{\mathrm{IIi}}^{\mathrm{T}}\boldsymbol{\beta}_{II} + u_{IIi} + e_{ij} \qquad u_{IIi} \sim N(0, \sigma_{u_{II}}^2), \qquad e_{ij} \sim N(0, \sigma_e^2)$$

The correlated model allows for $\mathrm{cov}(u_I, u_{II}) > 0$. This takes into account that those individuals who consumed a food most frequently tend to consume more of it.

Note:

- Both parts of the model are estimated simultaneously.
- The Box-Cox parameter (lambda) and the covariate effects are estimated at the same time during the model fitting so that the best transformation is chosen after adjusting for these covariates.

5.2 NCI SAS Macros

The NCI method can be implemented with 3 SAS macros to model usual dietary intake for a single dietary component (whether consumed daily or episodically):

- **MIXTRAN**: Fits a nonlinear mixed model (using SAS PROC NLMIXED) on multiple 24-hour recalls to estimate the regression coefficients for mean usual intake.

  CAUTION: Do not use the standard errors (SE) and p-values calculated in this macro for analysis of HCHS/SOL data as they are only valid for data collected via Simple Random Sampling (SRS). For a complex survey design such as the one in HCHS/SOL they need to be calculated using replication methods of variance estimation (e.g. jackknife, bootstrap, or Balanced Repeated Replication (BRR) (Korn & Graubard, 1999) with pseudo-PSU to have two PSU per strata).

- **DISTRIB:** Estimates the distribution of usual intake of a food or nutrient in a population from the estimated parameters of the model specified in MIXTRAN. This macro simulates a population that has the same characteristics (as described by the values of the covariates) as the sample used to fit the model. It uses the Monte Carlo method to empirically estimate the distribution (mean and percentiles) of usual intake.

- **INDIVINT:** Predicts individual food or nutrient intake.

To estimate the distribution of usual intake you only need to execute MIXTRAN and DISTRIB macros, whereas to predict usual intake you need MIXTRAN and INDIVINT macros. Macros, User's Guide and Sample Programs (NHANES data) can be downloaded from:
http://riskfactor.cancer.gov/diet/usualintakes/macros_single.html

5.3 Q & A

5.3.1 What is the difference between ESTIMATING and PREDICTING usual intake?

In the Frequentist framework of statistics, we ESTIMATE parameters (unknown and fixed) and we PREDICT random variables. If we want to describe the **usual dietary intake of a population** or subpopulation we **ESTIMATE** parameters of interest (mean, percentiles, etc.). If we want to study the association between usual dietary intake and an outcome then we follow three steps: (1) estimate mean parameters, (2) **PREDICT dietary intake for each specific individual**, and (3) use these values to assess the association of interest.

5.3.2 Why do we need to use the NCI method to estimate usual intake?

The NCI method:

- estimates within-person (i.e. day-to-day) and between-person variability. It models repeated 24-hour dietary recalls to account for the large day-to-day variability within persons.
- accommodates (through a Box-Cox transformation) **highly skewed consumption amounts.**
- allows correlated errors for probability of consumption and amount consumed to be correlated.
- **relates covariates to usual intake**, which in turn
    - helps improve the estimates by explaining the variability (e.g. differences due to weekend, sequence, mode, etc.) and
    - allows subpopulation analyses. For example, it allows for comparison of median intake of calcium between Hispanic/Latino background.
    - Note that the parameter estimates for the covariates are on the transformed scale. Hence, when they are back transformed for interpretation purposes, the relative order is maintained. Because for a symmetric distribution the mean and the median are the same the backtransformation is to the median, not to the mean on the original scale, where the distribution is skewed.
- uses Monte Carlo method in the DISTRIB macro which estimates the usual intake distribution (i.e. percentiles) in addition to the mean.
- aggregates consumption and amount for usual intake predictions.
- when using the FPQ as a covariate it can substantially improve the power to detect relationships between dietary intakes as predictor variables and outcomes. The magnitude of improvement depends on the proportion of zeros in the report of the dietary component, with the FPQ having a great impact for those with a large number of zero intakes.
- See NCI Measurement Error Webinars 1, 2 and 3 for details.

5.3.3 Why do we need covariates to estimate usual intake?

In nutritional epidemiology we are very often interested in subpopulation analyses. This allows us to 1) monitor subpopulations (e.g. Hispanic/Latino background, pregnant women, etc.), and 2) identify personal characteristics associated with usual intake (e.g. smoking status, education, etc.). These characteristics help better predict usual intake for each individual.

5.3.4 When do I need to use the Food Propensity Questionnaire (FPQ)?

The strength of the FPQ is to improve estimates for episodically consumed foods or food-groups. The FPQ assess long-term dietary intake and captures 'truly' non-consumers. In contrast, 24-hour recalls are short-term instruments and would overestimate the percent of non-consumers because of the large day-to-day variability within persons.

- We recommend using the FPQ when estimating food group usual intake, but note that incorporating FPQ data can limit the sample to those who completed the FPQ.

- You DO NOT need the FPQ to estimate or predict nutrient usual intake. However, there is some evidence that it could help improve precision of the estimates (Kipnis et al, *Biometrics* 2009).

- See NCI Measurement Error Webinars: 1, 3, 8 and 10.

5.3.5 Other questions and answers

See Module 18 "Modeling Usual Intake using Dietary Recall Data" from the NCI webinar for other questions such as:

- How does the NCI method adjust for weekend and weekday consumption?
- What are the assumptions of the NCI method?
- What important caveats are associated with the NCI Method?

6 Model Used to Estimate Usual Nutrient Intake in HCHS/SOL

Most nutrients are consumed every day so usual intake was modeled using an amount model (i.e. one-part nonlinear mixed model), except for alcohol which was modeled as a two-part model. The NCI method classifies covariates as individual, time dependent and nuisance:

1. Individual level (affects true intake on all days): gender, age, field center and Hispanic/Latino background

2. Time dependent (affects true intake on specific days): weekend and self-report intake amount (more or less than usual amount)

3. Nuisance (affects reporting error): sequence/mode (1st recall is in-person and 2nd recall is by telephone)

Mathematically the model is specified as,

$$g(R_{ij},\lambda) = \beta_0 + \beta_1 bkgr1_i + \ldots + \beta_6 bkgr6_i + \beta_7 age_i + \beta_8 male_i + \beta_9 weekend_i + \beta_{10} 2ndrecall_i + \beta_{11} more_i + \beta_{12} less_i + \beta_{13} center1_i + \beta_{14} center2_i + \beta_{15} center3_i + u_i + e_{ij}$$

- **Bkgrd1, Bkgrd2, …, Bkgrd6:** Indicator variables for Hispanic background. We recommend using Mexican as the reference group.
- **Center1, Center2, and Center3:** Indicator variables for center.
- **Weekend:** Indicator variable for weekend which is defined as Friday, Saturday and Sunday from date of intake (variable DTSA3).

- **2nd recall:** Indicator variable for second 24-hour recall (mostly telephone administered) from variable recallnum.
  1 = First 24-hour recall
  2 = Second 24-hour recall
- **More and less**: Two dummy variables from self-reported intake amount (variable DTIA15):
  0 = Close to the amount that you usually eat?
  1 = A lot more than usual?
  2 = A lot less than usual?

How does the NCI method estimate the distribution of population nutrient intake in HCHS/SOL?

1. **Individual level covariates.** The simulated population has the same covariate pattern as the HCHS/SOL sample.

2. **Time-dependent covariates**. Since we want a single usual intake estimate for each person, the simulated population sets DTIA15=0 (i.e. more=less=0). Similarly, the simulated population uses weighted average of weekday and weekend day (weights of 4/7 and 3/7, respectively) to aggregate over all days of the week in proportion to how often occur.

3. **Nuisance parameters** (sequence/mode). The distribution is estimated for the first recall only; it is adjusted for the 2nd recall (as for other covariates included in the model).

7 Estimated Usual Food-group Intake (HCHS/SOL Manuscript #9 Usual Dietary Intake)

Current HCHS/SOL dietary release does not include predicted usual food group intake.
Manuscript #9 "Usual dietary intake in HCHS/SOL" by Siega-Riz AM et al estimated usual food group intake by the NCI method at the population level, it did not predict food group intake for HCHS/SOL participants. A two-part model was used for most food groups. Vegetables (all and other), grains (refined and all), meat, milk, diet beverages, sugar (all) and fat (from oils and all) were estimated using an amount-only model because these foods were almost universally consumed by everyone. The first part of the model estimated the probability of consumption using logistic regression with a person-specific random effect and the second part specified the consumption-day amount using linear regression on a transformed scale, also with a person-specific effect. The person-specific random effects were allowed to be correlated across the two parts, because the probability of consumption is often related to the amount consumed (Subar, J Am Diet Assoc 2006). The same covariates were specified in both parts of the model, and the corresponding food group from the FPQ was included as a covariate to improve estimates for episodically consumed foods. Repeated 24hr recalls captured the natural day-to-day (within-individual) variation in dietary intake. The FPQ captured the consumption of episodically consumed foods. This information can substantially improve the power to detect associations between dietary intakes as predictor variables and health outcomes, especially for foods that were not commonly consumed.

8 HCHS/SOL Recommendations for Diet Data Use

- It is highly recommended that HCHS/SOL investigators include a member of the Diet and Supplements Subcommittee or another competent nutritionist familiar with NDSR output and nutrition exposures as a part of their analysis and writing teams. This will help standardize the use of nutrition-related variables and the interpretation of results.

- Unreliable recalls according to the interviewer (DTIA16 = 1 or 2) are recommended to be excluded from analyses.

- Data cleaning should be done separately for each dietary recall since intake from 1st recalls is known to be higher than intake from 2nd recalls. Also, consider using self-reported intake amount (DTIA15) and NDSR "Notes from the Trailer tab" or "Food detail Window Notes" (e.g. variables DTIA154 , NTIA154 and DIEA141) to understand, determine and clean extreme low or high values. DTIA15, self-reported usual/higher/lower amount, may help validate extreme values.

9 RESOURCES

NDSR (Nutrition Data System for Research)
http://www.ncc.umn.edu/index.html)

Usual Dietary Intake
http://riskfactor.cancer.gov/diet/usualintakes/

Measurement error in dietary data (NCI Webinar Series)
http://riskfactor.cancer.gov/measurementerror/

The goals of the Webinar Series are to understand:

- the sources and magnitudes of dietary measurement errors;
- how measurement error may affect estimates of usual dietary intake distributions;
- how measurement error may affect analyses of diet-health relationships;
- how the effects of measurement error may be mitigated.

It is organized by collaborators from NCI, Office of Dietary Supplements, USDA, Gertner Institute, Texas A&M University, and Wake Forest University. The series is intended for nutritionists, epidemiologists, statisticians, graduate students, and others with an interest in measurement error in dietary intake data. Archived webinars are available at: http://riskfactor.cancer.gov/measurementerror/. An intermediate level of familiarity with statistics and dietary assessment is recommended. Webinars of particular interest for section 5 are:

- Webinar #2 "Estimating usual intake distributions for dietary components consumed daily by nearly all persons" describes statistical modeling techniques and data requirements for estimating usual intake.

- Webinar #6 "The problem of measurement error when examining diet-health relationships" explains types and magnitude of measurement error that occur in dietary data, statistical models

for evaluating diet-health relationships (including energy adjustment models), and the qualitative and quantitative impact of measurement error on studies of diet-health relationships.

Advanced Dietary Analyses (NHANES)
http://www.cdc.gov/nchs/tutorials/dietary/advanced/index.htm

Other methods to estimate or predict usual intake include:

Iowa State University Method for Estimation of Usual Intake (ISU and ISUF)
http://streaming.stat.iastate.edu/cssm/index.php?option=com_content&view=article&id=38&Itemid=73

The Multiple Source Method (MSM)
https://msm.dife.de/

10 REFERENCES

Dekkers ALM, Verkaik J, Van Rossum CTM, Slob W, Ocké MC.*SPADE: Statistical Program To Assess Dietary Exposure—User's Manual*. National Institute for Public Health and Environment: Bilthoven (in preparation).

Dodd KW, Guenther PM, Freedman LS, Subar AF, Kipnis V, Midthune D, Tooze JA, Krebs-Smith SM. Statistical methods for estimating usual intake of nutrients and foods: a review of the theory. *J Am Diet Assoc* 2006 Oct;106(10):1640-50.

Kipnis V, Midthune D, Buckman DW, Dodd KW, Guenther PM, Krebs-Smith SM, Subar AF, Tooze JA, Carroll RJ, Freedman LS. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics* 2009 Dec;65(4):1003-10.

Korn EL and Graubard BI (1999). Analysis of Health Surveys. Wiley Inter-Science.

Subar AF, Dodd KW, Guenther PM, Kipnis V, Midthune D, McDowell M, Tooze JA, Freedman LS, Krebs-Smith SM. The food propensity questionnaire: concept, development, and validation for use as a covariate in a model to estimate usual food intake. *J Am Diet Assoc* 2006 Oct;106(10):1556-63.

Tooze JA, Midthune D, Dodd KW, Freedman LS, Krebs-Smith SM, Subar AF, Guenther PM, Carroll RJ, Kipnis V. A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J Am Diet Assoc* 2006 Oct;106(10):1575-87.

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 Nov 30;29(27):2857-68.

Parson R, Munuo SS, Buckman DW, Tooze JA, and Dodd KW, 2009. User's Guide of Analysis of Usual Intake accessed from http://riskfactor.cancer.gov/diet/usualintakes/macros_single.html

NOTE: For more references, see 'Recommended Resources' at NCI Measurement Error Webinar Series

HCHS/SOL Manuscripts

#8 "Dietary assessment methodology in diverse Latino groups: Hispanic Community Health Study/Study of Latinos" by Himes J and (in alphabetical order) Ayala, GX, Catellier, DJ, Daviglus, ML, Gellman, MD, Isasi, CR, Loria, CM, Mossavar-Rahmani, Y, Rock, CL, Shor-Posner, G, Siega-Riz, AM, Van Horn, L

#9 "Usual dietary intake in HCHS/SOL" by Siega-Riz AM and (in alphabetical order) Ayala GX, Gellman MD, Ginsberg M, Himes JH, Liu K, Loria CM, Mossavar-Rahmani Y, Rock CL, Rodriguez B, Sotres-Alvarez D, Van Horn L.

#39 "Reduction of elements in a food frequency questionnaire for estimation of usual food intake: An example from the HCHS/SOL" by Catellier DJ and (in alphabetical order) Ayala GX, Gellman MD, Himes JH, Isasi CR, Liu K, Mossavar-Rahmani Y, Ou FS, Siega-Riz AM, Sotres-Alvarez D, Van Horn, L

APPENDIX A. Food Group Derived Variables (dietary recall level)

For information on NDSR Food Group Serving Count System and serving sizes see Appendix 10 of the NDSR 2011 Manual.

Table A. Food Group Derived Variables from 24hr Dietary Recalls

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
FRUIT_CIT	Fruit – Citrus, servings/day	DTSA4	FRU0100 Citrus Juice
		DTSA6	FRU0300 Citrus Fruit
FRUIT_OTH	Fruit – Others, servings/day	DTSA5	FRU0200 Fruit Juice excluding Citrus Juice
		DTSA7	FRU0400 Fruit excluding Citrus Fruit
		DTSA8	FRU0500 Avocado and Similar
		DTSA9	FRU0600 Fried Fruits
		DTSA10	FRU0700 Fruit-based Savory Snack
FRUIT_ALL	Fruit – Others, servings/day	FRUIT_CIT	Fruit – Citrus, servings/day
		FRUIT_OTH	Fruit – Others, servings/day
FRUIT_ALL_WOJUICE	Fruit – Overall, without fruit juice, servings/day	DTSA6	FRU0300 Citrus Fruit
		DTSA7	FRU0400 Fruit excluding Citrus Fruit
		DTSA8	FRU0500 Avocado and Similar
VEG_DARK	Vegetable – Dark-green, servings/day	DTSA11	VEG0100 Dark-green Vegetables
VEG_ORAN	Vegetable – Orange, servings/day	DTSA12	VEG0200 Deep-yellow Vegetables
VEG_TOMA	Vegetable – Tomato, servings/day	DTSA13	VEG0300 Tomato
VEG_WPOT	Vegetable – White Potatoes, servings/day	DTSA14	VEG0400 White Potatoes
VEG_STAR	Vegetable – Starchy, servings/day	DTSA15	VEG0800 Fried Potatoes
		DTSA16	VEG0450 Other Starchy Vegetables
VEG_BEAN	Vegetable – Beans, servings/day	DTSA17	VEG0700 Legumes (cooked dried beans)
VEG_OTH	Vegetable - Other, servings/day	DTSA18	VEG0600 Other Vegetables
		DTSA19	VEG0900 Fried Vegetables
		DTSA20	VEG0500 Vegetable Juice
		DTSA164	MSC0500 Pickled Foods
VEG_ALL	Vegetable – Overall, servings/day	VEG_DARK	Vegetable – Dark-green, servings/day
		VEG_ORAN	Vegetable – Orange, servings/day
		VEG_TOMA	Vegetable – Tomato, servings/day
		VEG_WPOT	Vegetable – White Potatoes, servings/day
		VEG_STAR	Vegetable – Starchy, servings/day

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
		VEG_BEAN	Vegetable – Beans, servings/day
		VEG_OTH	Vegetable - Other, servings/day
VEG_ALL_WOPOTATO	Vegetable – Overall, without potato , servings/day	DTSA11	VEG0100 Dark-green Vegetables
		DTSA12	VEG0200 Deep-yellow Vegetables
		DTSA13	VEG0300 Tomato
		DTSA16	VEG0450 Other Starchy Vegetables
		DTSA18	VEG0600 Other Vegetables
		DTSA20	VEG0500 Vegetable Juice
VEG_ALL_WOFPOTATO	Vegetable – Overall, without fried potato, servings/day	DTSA11	VEG0100 Dark-green Vegetables
		DTSA12	VEG0200 Deep-yellow Vegetables
		DTSA13	VEG0300 Tomato
		DTSA14	VEG0400 White Potatoes
		DTSA16	VEG0450 Other Starchy Vegetables
		DTSA17	VEG0700 Legumes (cooked dried beans)
		DTSA18	VEG0600 Other Vegetables
		DTSA19	VEG0900 Fried Vegetables
		DTSA20	VEG0500 Vegetable Juice
		DTSA164	MSC0500 Pickled Foods
FRUITVEG_WOFPOTATO	Fruit and vegetables without fried potato (servings/day)	FRUIT_ALL	Fruit – Others, servings/day
		DTSA15	VEG0800 Fried Potatoes
		VEG_ALL	Vegetable – Overall, servings/day
GRAIN_REF	Grain – Refined Grain, servings/day	DTSA23	GRS0100 Grains, Flour and Dry Mixes - Some Whole Grain
		DTSA24	GRR0100 Grains, Flour and Dry Mixes - Refined Grain
		DTSA26	GRS0200 Loaf-type Bread and Plain Rolls - Some Whole Grain
		DTSA27	GRR0200 Loaf-type Bread and Plain Rolls - Refined Grain
		DTSA29	GRS0300 Other Breads (quick breads, corn muffins, tortillas) - Some Whole Grain
		DTSA30	GRR0300 Other Breads (quick breads, corn muffins, tortillas) - Refined Grain
		DTSA32	GRS0400 Crackers - Some Whole Grain
		DTSA33	GRR0400 Crackers - Refined Grain
		DTSA35	GRS0500 Pasta - Some Whole Grain
		DTSA36	GRR0500 Pasta - Refined Grain
		DTSA38	GRS0600 Ready-to-eat Cereal (not presweetened) - Some Whole Grain
		DTSA39	GRR0600 Ready-to-eat Cereal (not presweetened) - Refined Grain
		DTSA41	GRS0700 Ready-to-eat Cereal (presweetened) - Some Whole Grain
		DTSA42	GRR0700 Ready-to-eat Cereal (presweetened) - Refined Grain
		DTSA44	GRS0800 Cakes, Cookies, Pies, Pastries, Danish, Doughnuts and Cobblers - Some Whole Grain
		DTSA45	GRR0800 Cakes, Cookies, Pies, Pastries, Danish, Doughnuts and Cobblers - Refined Grain
		DTSA47	GRS1000 Snack Bars - Some Whole Grain

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
		DTSA48	GRR1000 Snack Bars - Refined Grain
		DTSA50	GRS0900 Snack Chips - Some Whole Grain
		DTSA51	GRR0900 Snack Chips - Refined Grain
		DTSA53	GRW1200 Flavored Popcorn
		DTSA54	GRO0100 Baby Food Grain Mixtures
GRAIN_WHL	Grain – Whole Grain, servings/day	DTSA22	GRW0100 Grains, Flour and Dry Mixes - Whole Grain
		DTSA25	GRW0200 Loaf-type Bread and Plain Rolls - Whole Grain
		DTSA28	GRW0300 Other Breads (quick breads, corn muffins, tortillas) - Whole Grain
		DTSA31	GRW0400 Crackers - Whole Grain
		DTSA34	GRW0500 Pasta - Whole Grain
		DTSA37	GRW0600 Ready-to-eat Cereal (not presweetened) - Whole Grain
		DTSA40	GRW0700 Ready-to-eat Cereal (presweetened) - Whole Grain
		DTSA43	GRS0700 Ready-to-eat Cereal (presweetened) - Some Whole Grain
		DTSA46	GRW1000 Snack Bars - Whole Grain
		DTSA49	GRW0900 Snack Chips - Whole Grain
		DTSA52	GRW1100 Popcorn
GRAIN_ALL	Grain – Overall, servings/day	GRAIN_WHL	Grain – Whole Grain, servings/day
		GRAIN_REF	Grain – Refined Grain, servings/day
MEAT_RED	Meat – Red Meat, servings/day	DTSA55	MRF0100 Beef
		DTSA56	MRL0100 Lean Beef
		DTSA57	MRF0200 Veal
		DTSA58	MRL0200 Lean Veal
		DTSA59	MRF0300 Lamb
		DTSA60	MRL0300 Lean Lamb
		DTSA61	MRF0400 Fresh Pork
		DTSA62	MRL0400 Lean Fresh Pork
		DTSA65	MRF0500 Game
MEAT_LUNCH	Meat – Luncheon, servings/day	DTSA63	MCF0200 Cured Pork
		DTSA64	MCL0200 Lean Cured Pork
		DTSA74	MCF0100 Cold Cuts and Sausage
		DTSA75	MCL0100 Lean Cold Cuts and Sausage
MEAT_POUL	Meat – Poultry, servings/day	DTSA66	MPF0100 Poultry
		DTSA67	MPL0100 Lean Poultry
		DTSA68	MPF0200 Fried Chicken - Commercial Entrée and Fast Food
MEAT_FISH	Meat – Fish, servings/day	DTSA69	MFF0100 Fish - Fresh and Smoked
		DTSA70	MFL0100 Lean Fish - Fresh and Smoked
		DTSA71	MFF0200 Fried Fish - Commercial Entrée and Fast Food
		DTSA72	MSL0100 Shellfish

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
		DTSA73	MSF0100 Fried Shellfish - Commercial Entrée and Fast Food
MEAT_ORG	Meat – Organ Meats, servings/day	DTSA76	MOF0100 Organ Meats
MEAT_EGG	Meat – Eggs, servings/day	DTSA79	MOF0300 Eggs
		DTSA80	MOF0400 Egg Substitute
MEAT_NUT	Meat – Nuts, servings/day	DTSA81	MOF0500 Nuts and Seeds
		DTSA82	MOF0600 Nut and Seed Butters
MEAT_SOY	Meat – Soy, servings/day	DTSA83	MOF0700 Meat Alternatives
		DTSA117	DOT0800 Infant Formula - Nondairy
MEAT_ALL	Meat – Overall (servings/day)	MEAT_RED	Meat – Red Meat, servings/day
		MEAT_LUNCH	Meat – Luncheon, servings/day
		MEAT_POUL	Meat – Poultry, servings/day
		MEAT_FISH	Meat – Fish, servings/day
		MEAT_ORG	Meat – Organ Meats, servings/day
		MEAT_EGG	Meat – Eggs, servings/day
		MEAT_NUT	Meat – Nuts, servings/day
		MEAT_SOY	Meat – Soy, servings/day
NUT_LEGUMES	Nut and Legumes, servings/day	DTSA17	VEG0700 Legumes (cooked dried beans)
		DTSA81	MOF0500 Nuts and Seeds
		DTSA82	MOF0600 Nut and Seed Butters
MILK_MILK	Milk – Milk, servings/day	DTSA84	DMF0100 Milk - Whole
		DTSA85	DMR0100 Milk - Reduced Fat
		DTSA86	DML0100 Milk - Low Fat and Fat Free
		DTSA87	DMN0100 Milk - Nondairy
		DTSA88	DMF0200 Ready-to-drink Flavored Milk - Whole
		DTSA89	DMR0200 Ready-to-drink Flavored Milk - Reduced Fat
		DTSA90	DML0200 Ready-to-drink Flavored Milk - Low Fat and Fat Free
		DTSA91	DML0300 Sweetened Flavored Milk Beverage Powder with Non-fat Dry Milk
		DTSA92	DML0400 Artificially Sweetened Flavored Milk Beverage Powder with Non-fat Dry Milk
		DTSA93	SWT0600 Sweetened Flavored Milk Beverage Powder without Non-fat Dry Milk
		DTSA94	MSC1100 Artificially Sweetened Flavored Milk Beverage Powder without Non-fat Dry Milk
		DTSA114	DOT0500 Dairy-based Sweetened Meal Replacement/Supplement
		DTSA115	DOT0600 Dairy-based Artificially Sweetened Meal Replacement/Supplement
		DTSA116	DOT0700 Infant Formula
MILK_CHES	Milk – Cheese, servings/day	DTSA95	DCF0100 Cheese - Full Fat
		DTSA96	DCR0100 Cheese - Reduced Fat
		DTSA97	DCL0100 Cheese - Low Fat and Fat Free

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
		DTSA98	DCN0100 Cheese - Nondairy
MILK_YOGU	Milk – Yogurt, servings/day	DTSA99	DYF0100 Yogurt - Sweetened Whole Milk
		DTSA100	DYR0100 Yogurt - Sweetened Low Fat
		DTSA101	DYL0100 Yogurt - Sweetened Fat Free
		DTSA102	DYF0200 Yogurt - Artificially Sweetened Whole Milk
		DTSA103	DYR0200 Yogurt - Artificially Sweetened Low Fat
		DTSA104	DYL0200 Yogurt - Artificially Sweetened Fat Free
		DTSA105	DYN0100 Yogurt - Nondairy
MILK_DESR	Milk – Dessert, servings/day	DTSA106	DOT0100 Frozen Dairy Dessert
		DTSA108	DOT0300 Pudding and Other Dairy Dessert
		DTSA109	DOT0400 Artificially Sweetened Pudding and Other Dairy Dessert
MILK_ALL	Milk – Overall, servings/day	MILK_MILK	Milk – Milk, servings/day
		MILK_CHES	Milk – Cheese, servings/day
		MILK_YOGU	Milk – Yogurt, servings/day
		MILK_DESR	Milk – Dessert, servings/day
FAT_DISC	Fat – Discretionary Fat, servings/day	DTSA110	FCF0100 Cream
		DTSA111	FCR0100 Cream - Reduced Fat
		DTSA112	FCL0100 Cream - Low Fat and Fat Free
		DTSA113	FCN0100 Cream - Nondairy
FAT_OIL	Fat – Oil, servings/day	DTSA118	FMF0100 Margarine - Regular
		DTSA119	FMR0100 Margarine - Reduced Fat
		DTSA120	FOF0100 Oil
		DTSA121	FSF0100 Shortening
		DTSA122	FAF0100 Butter and Other Animal Fats - Regular
		DTSA123	FAR0100 Butter and Other Animal Fats - Reduced Fat
		DTSA124	FDF0100 Salad Dressing - Regular
		DTSA125	FDR0100 Salad Dressing - Reduced Fat/Reduced Calorie/Fat Free
		DTSA160	MSC0100 Gravy - Regular
		DTSA161	MSC0200 Gravy - Reduced Fat/Fat Free
		DTSA162	MSC0300 Sauces and Condiments - Regular
		DTSA163	MSC0400 Sauces and Condiments - Reduced Fat
FAT_ALL	Fat – Overall, servings/day	FAT_DISC	Fat – Discretionary Fat, servings/day
		FAT_OIL	Fat – Oil, servings/day
SUGAR_SGR	Sugar – Sugar, servings/day	DTSA126	SWT0400 Sugar
		DTSA127	MSC1200 Sugar Substitute
		DTSA128	SWT0500 Syrup, Honey, Jam, Jelly, Preserves
		DTSA129	SWT0700 Sauces, Sweet - Regular
		DTSA130	SWT0800 Sauces, Sweet - Reduced Fat/Reduced Calorie/Fat Free

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
SUGAR_DESR	Sugar – Dessert, servings/day	DTSA131	SWT0100 Chocolate Candy
		DTSA132	SWT0200 Non-chocolate Candy
		DTSA133	SWT0300 Frosting or Glaze
		DTSA165	MSC0600 Miscellaneous Dessert
SUGAR_SWTB	Sugar – Sugar Sweetened Beverage, servings/day	DTSA134	BVS0400 Sweetened Soft Drinks
		DTSA137	BVS0300 Sweetened Fruit Drinks
		DTSA139	BVS0500 Sweetened Tea
		DTSA142	BVS0100 Sweetened Coffee
		DTSA145	BVS0200 Sweetened Coffee Substitutes
		DTSA148	BVS0600 Sweetened Water
		DTSA151	BVS0700 Nondairy-based Sweetened Meal Replacement/Supplement
		DTSA154	BVO0100 Non-alcoholic Beer
		DTSA155	BVO0200 Non-alcoholic Light Beer
SUGAR_SWTB _WJUICE	Sugar Sweetened Beverage, include fruit juice, servings/day	DTSA4	FRU0100 Citrus Juice
		DTSA5	FRU0200 Fruit Juice excluding Citrus Juice
		DTSA134	BVS0400 Sweetened Soft Drinks
		DTSA137	BVS0300 Sweetened Fruit Drinks
		DTSA139	BVS0500 Sweetened Tea
		DTSA142	BVS0100 Sweetened Coffee
		DTSA145	BVS0200 Sweetened Coffee Substitutes
		DTSA148	BVS0600 Sweetened Water
		DTSA151	BVS0700 Nondairy-based Sweetened Meal Replacement/Supplement
		DTSA154	BVO0100 Non-alcoholic Beer
		DTSA155	BVO0200 Non-alcoholic Light Beer
SUGAR_DIETB	Sugar – Diet Beverage, servings/day	DTSA135	BVA0400 Artificially Sweetened Soft Drinks
		DTSA136	BVU0300 Unsweetened Soft Drinks
		DTSA138	BVA0300 Artificially Sweetened Fruit Drinks
		DTSA140	BVA0500 Artificially Sweetened Tea
		DTSA141	BVU0400 Unsweetened Tea
		DTSA143	BVA0100 Artificially Sweetened Coffee
		DTSA144	BVU0100 Unsweetened Coffee
		DTSA146	BVA0200 Artificially Sweetened Coffee Substitutes
		DTSA147	BVU0200 Unsweetened Coffee Substitutes
		DTSA149	BVA0600 Artificially Sweetened Water
		DTSA152	BVA0700 Nondairy-based Artificially Sweetened Meal Replacement/Supplement
		DTSA153	BVU0600 Nondairy-based Unsweetened Meal Replacement/Supplement
SUGAR_ALL	Sugar – Overall, servings/day	SUGAR_SGR	Sugar – Sugar, servings/day
		SUGAR_DESR	Sugar – Dessert, servings/day

Food Group (Variable name)	Label	HCHS/SOL source variable name	Source Variable Description
		SUGAR_SWTB	Sugar – Sugar Sweetened Beverage, servings/day
		SUGAR_DIETB	Sugar – Diet Beverage, servings/day
WATER	Water, servings/day	DTSA150	BVU0500 Unsweetened Water
ALCOHOL	Alcohol, servings/day	DTSA156	BVE0100 Beer and Ale
		DTSA157	BVE0400 Cordial and Liqueur
		DTSA158	BVE0300 Distilled Liquor
		DTSA159	BVE0200 Wine
SNACK_SWT	Snack – Sweet Snack, servings/day	DTSA43	GRW0800 Cakes, Cookies, Pies, Pastries, Danish, Doughnuts and Cobblers - Whole Grain
		DTSA44	GRS0800 Cakes, Cookies, Pies, Pastries, Danish, Doughnuts and Cobblers - Some Whole Grain
		DTSA45	GRR0800 Cakes, Cookies, Pies, Pastries, Danish, Doughnuts and Cobblers - Refined Grain
		DTSA46	GRW1000 Snack Bars - Whole Grain
		DTSA47	GRS1000 Snack Bars - Some Whole Grain
		DTSA48	GRR1000 Snack Bars - Refined Grain
SNACK_SALT	Snack – Salty Snack, servings/day	DTSA49	GRW0900 Snack Chips - Whole Grain
		DTSA50	GRS0900 Snack Chips - Some Whole Grain
		DTSA51	GRR0900 Snack Chips - Refined Grain
		DTSA52	GRW1100 Popcorn
		DTSA53	GRW1200 Flavored Popcorn
SNACK_NUT	Snack – Nuts, servings/day	DTSA81	MOF0500 Nuts and Seeds
		DTSA82	MOF0600 Nut and Seed Butters
SNACK_CRACK	Snack – Cracker, servings/day	DTSA31	GRW0400 Crackers - Whole Grain
		DTSA32	GRS0400 Crackers - Some Whole Grain
		DTSA33	GRR0400 Crackers - Refined Grain
SNACK_VEGFR	Snack – Fruit or Vegetable Savory(servings/day)	DTSA10	FRU0700 Fruit-based Savory Snack
		DTSA21	FMC0100 Vegetable-based Savory Snack
SNACK_ALL	Snack – Overall, servings/day	SNACK_SWT	Snack – Sweet Snack, servings/day
		SNACK_SALT	Snack – Salty Snack, servings/day
		SNACK_NUT	Snack – Nuts, servings/day
		SNACK_CRACK	Snack – Cracker, servings/day
		SNACK_VEGFR	Snack – Fruit or Vegetable Savory(servings/day)

APPENDIX B. Data cleaning for 24-hr recalls based on daily energy intake

There is not a single best way to clean diet data for extreme values. Data cleaning methods depend on the diet instrument used to assess dietary intake, the specific nutrient or food of interest, and the study population, among other things. We suggest the investigator or data analyst consult with someone who has experience with analyzing dietary data. In particular, for 24hr dietary recalls it is recommended that data cleaning is done separately for each dietary recall since intake from 1st recall is known to be higher than intake from the 2nd recall. We have found this to be the case with in HCHS/SOL as well.

Also, consider using self-reported intake amount (DTIA15) and NDSR "Notes from the Trailer tab" or "Food detail Window Notes" (e.g. variables DTIA154 , NTIA154 and DIEA141) to understand, determine and clean extreme low or high values. Examples include two participants with zero energy intake due to fasting (identified by DTIA154), and extreme values confirmed (e.g. "14 chicken wings eaten. Confirmed by participant" identified by DIEA141), etc.

HCHS/SOL diet derived variables distributed to investigators (FOOD_GROUPS_DERV, PRED_NUTR_DERV and derived diet scores in PART_DERV) excluded dietary recalls based on the following 3 criteria; there was no other data cleaning done specific to each nutrient or food group.

1. unreliable according to the interviewer (DTIA16)
2. had extreme observed daily energy intake (DTIA20) defined with the sequence-gender specific unweighted 1st percentile or above the 99th percentile calculated from reliable recalls (DTIA16).

Table B. First and 99th percentiles of energy intake (kcal) by gender and recall

Gender	Recall	1st percentile	99th percentile
Female	1st	425.83	4,014.53
Female	2nd	356.71	3,465.15
Male	1st	607.76	6,155.40
Male	2nd	511.19	4,912.40

3. had extreme energy intake at the food level (DIEA8) which, upon case investigation was deemed to be a data entry error.

Below is SAS code to exclude dietary recalls according to those 3 criteria.

```
/* GLOBAL INITIALIZATION */

%let prog   = ABC;                          ** programmer          **;
%let job    = Appendix_B_DietDataCleaning;  ** job name (SAS code) **;
options nodate nonumber;
footnote "Job &job run by &prog on &SYSDATE at &SYSTIME";

* HCHS DATASETS;
libname main "H:\HCHS\";

* SPECIFY THE DIRECTORY WHERE THE OUTPUT DATASET WILL BE STORED;
libname outlib 'H:\HCHS\OUTPUT';
```

```
data dtia;
   merge main.dtia (in=indtia   keep = ID recallnum dtia3 dtia15 dtia16 dtia20)
         main.part_derv (in=partderv keep = ID gendernum);
   by ID;
   if first.ID then do;
      N_RECALLS_AVAILABLEORIG=0;
   end;

   if recallnum>.Z then N_RECALLS_AVAILABLEORIG+1;
   label
 N_RECALLS_AVAILABLEORIG='Number of 24hr recalls originally available per subject';
 run;

 proc means data = dtia p1 p99 nway noprint;
      var dtia20;
      class gendernum recallnum;
      output out=temp1 p1=p1 p99=p99;
      where dtia16=0;
     title "Percentiles 1 and 99 AMONG reliable recalls (DTIA16=0)";
 run;

 proc print data=temp1;
   title 'Percentiles 1 and 99 by gender and recall AMONG reliable (DTIA16=0)';
 run;

 proc sort data=dtia; by gendernum recallnum; run;

 data OUTLIB.dtia_cleaned (label = "DTIA cleaned created in job &job"
                        drop = p1 p99 flag_extr1 flag_extr99 flag_extr gendernum);
   merge dtia&rt_frz(in=indtia)
         temp1(keep=gendernum recallnum p1 p99);
   by gendernum recallnum;

   if dtia20<p1  then flag_extr1=1;
   if dtia20>p99 then flag_extr99=1;

/* HARD CODES because data entry errors (identified by DIEA8 (energy at food level)
being extremes but their overall energy intake (DTIA20 is not)
Per Request. */
   if id='33427131' & recallnum=1 then flag_extr99=1;
   if id='16720544' & recallnum=1 then flag_extr99=1;
   if id='30252101' & recallnum=2 then flag_extr99=1;

   if flag_extr1=1 | flag_extr99=1 then flag_extr=1;
   else flag_extr=0;

   if flag_extr=0 & dtia16=0;
 run;

proc freq data=OUTLIB.dtia_cleaned;
 table recallnum dtia16;
 title "DATA CHECK: CLEANED RECALLS";
run; title;

/* End of SAS code */
```

APPENDIX C. SAS code to Predict Energy Usual Intake Using NCI Macros

In this section, we use energy (variable name: DTIA20) as an example to illustrate how to execute NCI macros to obtain predicted energy usual intake. This code is an adaptation to HCHS/SOL data based on examples 1 and 4 from the "User's Guide of Analysis of Usual Intake" (Parson et al, 2009). Macros, User's Guide and example code (NHANES data) can be downloaded from: http://riskfactor.cancer.gov/diet/usualintakes/macros_single.html

There are two SAS macros needed to predict usual dietary intake for a single dietary component (whether consumed daily or episodically): **MIXTRAN** and **INDIVINT.** The MIXTRAN macro estimates the parameters, while the INDIVINT macro predicts usual nutrient intake for HCHS/SOL participants. We divide this task in four steps, of which the last three include SAS code.

Step 1: Specify the model

To estimate the usual energy intake distribution with the NCI method, we estimate the within and between person variance components and correct for the excessive intra-individual variation intrinsic of 24hr recalls (i.e. individuals do not eat the same foods and amounts every day). Recalls with daily energy intake (DTIA20) below the sequence-gender specific 1st percentile or above the 99th percentile or unreliable according to the interviewer (DTIA16) were excluded. See appendix B for details in data cleaning. Usual energy intake is estimated with a one-part nonlinear mixed model adjusted for gender, age, Hispanic/Latino background, site, weekend (including Friday), self-report intake amount (more, same or less than usual amount), and sequence (1st recall or 2nd recall). **Participants with at least one dietary recall are included in the model.** We include Hispanic/Latino background as a covariate in the model, and by specifying it in the subgroup option the MIXTRAN MACRO estimates the associated parameters (i.e. provides a separate usual intake distribution for each subgroup). **Note that only the covariate values differ between background groups; all other variance components remain the same for all backgrounds.**

Step 2: Data management

First, we merge covariates from baseline data with 24hr recall data. The PART_DERV dataset has covariates for the model (age, gender, Hispanic/Latino background) and study design variables (WEIGHT_FINAL_EXPANDED, STRAT, PSU_ID). The DTIA dataset contains nutrients from both 24hr recalls (e.g. DTIA20 for energy) with one record per recall; ID and RECALLNUM (1st or 2nd recall) uniquely identify a record for this dataset. Hence, there are at most two records per participant. The analysis dataset is called SOL.

```
* HCHS IS WHERE THE DATASETS ARE STORED;
libname hchs "H:\HCHS";

data part_derv;
    set hchs.part_derv(keep=id weight_final_expanded age gender strat psu_id
    center bkgrd1_c7);
    if gender ne " " then male=(gender="M");
    label male = "Male vs. Female";
run;
```

```
data dtia_cleaned;
    set hchs.dtia_cleaned(keep = id recallnum dtia3 dtia15 dtia20);
run;

proc sort data = part_derv; by id; run;
proc sort data = dtia_cleaned; by id recallnum; run;

data SOL;
    merge part_derv dtia_cleaned;
    by id;
run;
```

NCI macros require:

- covariates to be dummy variables
- sequence (1st or 2nd recall) and weekend to be specified as parameters in the NCI macros (and not as covariates)
- the sampling weight variable to be an integer
- subject ID to be numeric
- the weekend variable to be named “weekend”
- (MIXTRAN only) a variable that indexes repeated observations (RECALLNUM) for each subject to be an integer value of 1 or more.

```
data SOL;
    set SOL;
    * THE NCI MACRO HAS A BUG, THE WEEKEND VARIABLE NEED TO BE NAMED AS WEEKEND;
    * OTHERWISE, THE MACRO WILL STOP EXECUTING;
    label weekend = "Recall done during Fri.-Sun.";
    if dtia3 > .z then weekend = (weekday(dtia3) in (1, 6, 7));

    * WEIGHT VARIABLE NEEDS TO BE INTEGER. HENCE, WE ENED TO USE EXPANDED WEIGHTS;
    if weight_final_expanded > .z then wtc = round(weight_final_expanded);

    * CREATE DUMMY VARIABLES FOR HISPANIC BACKGROUND;
       bkgrd0 = (bkgrd1_c7 = 0);
       bkgrd1 = (bkgrd1_c7 = 1);
       bkgrd2 = (bkgrd1_c7 = 2);
       bkgrd3 = (bkgrd1_c7 = 3);
       bkgrd4 = (bkgrd1_c7 = 4);
       bkgrd5 = (bkgrd1_c7 = 5);
       bkgrd6 = (bkgrd1_c7 not in (0,1,2,3,4,5)); /* Due to missing and .Q */
    label bkgrd0 = 'Dummy for background = Dominican'
          bkgrd1 = 'Dummy for background = Central American'
          bkgrd2 = 'Dummy for background = Cuban'
          bkgrd3 = 'Dummy for background = Mexican'
          bkgrd4 = 'Dummy for background = Puerto Rican'
          bkgrd5 = 'Dummy for background = South American'
          bkgrd6 = 'Dummy for background = Mixed/Other/MISSING'
              ;
```

```
    if center ne ' ' then do;
        center1 = (center='B');
        center2 = (center='C');
        center3 = (center='M');
        center4 = (center='S');
    end;
    label center1 = 'Dummy for center = Bronx'
          center2 = 'Dummy for center = Chicago'
          center3 = 'Dummy for center = Miami'
          center4 = 'Dummy for center = San Diego'
              ;

    if recallnum > .z then recallnum2 = (recallnum = 2);
    if DTIA15 > .z then do;
        DTIA15_1 = (DTIA15 = 1);
        DTIA15_2 = (DTIA15 = 2);
    end;
    label
     recallnum = "Recall sequence (1=1st, 2=2nd)" /*For REPEAT parameter*/
     recallnum2 = "Dummy for second recall (0=1st, 1=2nd)"/*For SEQ parameter*/
     DTIA15_1 = "Dummy for considerably more than usual intake"
     DTIA15_2 = "Dummy for considerably less than usual intake"
              ;
run;
```

NOTE: for nutrients or foods that are consumed ubiquitously but still have a few zeros it is recommended to replace zero values with the smallest positive 24hr recall value.

Step 3: Execute NCI MIXTRAN SAS Macro

Input dataset: SOL (vertical format: one record per dietary recall; ID and RECALLNUM uniquely identify a record)

Output datasets:

Output Dataset	Brief description
Etas_energy.sas7bdat	Contains character strings that are interpreted by the MIXTRAN to calculate fixed effects predicted values.
parmsf2 energy.sas7bdat	Parameter estimates output by the SAS NLMIXED procedure for the 'Amount Model" in a base run; can be used as an input for starting values for same model a re-run.
_param_unc_energy.sas7bdat	Contains parameter estimates for the uncorrelated model and amount-only in one record
_pred_unc_energy.sas7bdat	Contains the predicted FIXED effects (transformed values) for both weekday and weekend day. It has one record per person

The "energy" suffix is specified by the user as the FOODTYPE parameter when calling MIXTRAN. The latter two datasets are saved for use as input for the INDIVINT macro. MIXTRAN names these output datasets: "etas", "paramsf2, "param", "unc", "pred". See "NCI User's Guide for Analysis of Usual Intakes" for the meaning of these conventions and variable names.

```
* SPECIFY THE DIRECTORY WHERE THE OUTPUT DATASETS WILL BE STORED;
libname outlib 'H:\HCHS\OUTPUT';

* INCLUDE THE MIXTRAN MACRO;
%include "H:\HCHS\NCI_Macros\mixtran_macro_v1.1.sas" / nosource;

* MPRINT option displays the SAS statements that are generated by macro execution
which is useful for debugging macros;
options mprint;

/* Mandatory parameters in THE MIXTRAN macro illustrated here in bold */
%mixtran (data = SOL,
      response = DTIA20,
      foodtype = energy,
      subject = id,
      repeat = recallnum,
      covars_prob = ,
      covars_amt = age male bkgrd0 bkgrd1 bkgrd2 bkgrd4 bkgrd5 bkgrd6 center1
      center2 center3 DTIA15_1 DTIA15_2,
      outlib = outlib,
      modeltype = amount,
      lambda = ,
      replicate_var = wtc,
      seq = recallnum2,
      weekend = weekend,
      vargroup = ,
      numvargroups = ,
      subgroup = bkgrd1_c7,
      start_val1 = ,
      start_val2 = ,
      start_val3 = ,
      vcontrol = ,
      nloptions = qmax=61 update=dbfgs, /* See SAS documentation */
      titles = ,
      printlevel = 2);
```

Step 4: Execute the **NCI INDIVINT macro**

This code is an adaptation to HCHS/SOL data from Example 4 in the "User's Guide of Analysis of Usual Intake" (Parson et al, 2009).

Input dataset: PARAMSUB1REC.sas7bdat (horizontal format: both recalls in one record; ID uniquely identifies a record). It is created from the original intake information (HCHS/SOL datasets) and datasets produced by MIXTRAN Macro. Detailed steps for data preparation are described below.

Output dataset: The INDIVINT macro will produce a dataset in the work library called "_resdata. sas7bdat". In this dataset, the variable INDUSINT is the final predicted intake for each individual (Kipnis and Tooze, personal communication). Variable INDUSINT and ID are saved for future use.

Step 4.1: Data management

The input dataset for this macro needs to be one observation per subject, and has to contain the following variables:

- predicted value from the MIXTRAN macro (_pred_unc_energy.sas7bdat)
- parameter estimates from MIXTRAN macro (_param_unc_energy.sas7bdat)
- ID
- Observed nutrient variable (DTIA20) from 1st and 2nd recall. Because the input dataset needs to be one observation per subject, energy intake (DTIA20) should be provided as two separate variables (e.g. R1=energy intake from first recall and R2=energy intake from second recall).
- If the model specified the weekend option in the MIXTRAN Macro, then pred_unc_energy.sas7bdat dataset has two variables x2b2_0 and x2b2_1 which are the predicted intake for weekday and weekend, respectively. A new variable X2B2 needs to be created as the weighted average of x2b2_0 and x2b2_1 (X2B2 = (x2b2_0*4 + x2b2_1*3) / 7). See section 6 for rationale.
- Finally, observations with nutrient intake = 0 will be replaced by half of the minimum nutrient intake from both recalls since it is an amount model and assumes intake > 0.

```
* COMBINE DATASETS WITH PREDICTED VALUE AND PARAMETER ESTIMATES FROM MIXTRAN;
data parampred;
    if (_n_ = 1) then set outlib._param_unc_energy;
    set outlib._pred_unc_energy;
run;

* STRING OUT INTAKE FROM 2 RECALLS SO THE DATASET CONTAINS 1 OBSERVATION PER
SUBJECT (HORIZONTAL FORMAT);
data subj1rec;
    set SOL;
    by id;
    retain R1 R2;
    if first.id then do;
       r1=.; r2=.;
    end;
    * STRING OUT 1ST AND 2ND RECALL INFORMATION;
    if recallnum eq 1 then R1 = dtia20;
        else if recallnum eq 2 then R2 = dtia20;

    if last.id then output;
run;

* CALCULATE THE SMALLEST POSITIVE 24-HR RECALL VALUE;
proc univariate data=SOL noprint;
    where dtia20 > 0;
    var dtia20;
    output out=outmin_amt min=min_amt;
run;

* COMBINE PARAMETER ESTIMATES WITH HORIZONTAL INTAKE AMOUNT;
data parsubj1rec;
    merge parampred subj1rec(keep = id R1 R2);
    by id;
run;
```

```
data paramsubj1rec; * DATASET WILL BE THE INPUT DATASET FOR THE INDIVINT MACRO;
    if _n_=1 then set outmin_amt;
    set parsubj1rec;

      * REPLACE INTAKE=0 WITH HALF THE MINIMUM. THIS IS NOT NECESSARY FOR ENERGY
      SINCE THERE ARE NO VALUES = 0 AFTER CLEANING DATA. BUT THIS IS NEEDED FOR 1-
      PART MODELS THAT DO HAVE FEW ZEROES SINCE THE MODEL ASSUMES INTAKE > 0;
    if R1 = 0 then R1 = min_amt*0.5;
    if R2 = 0 then R2 = min_amt*0.5;

    * WEIGHTED AVERAGE OF WEEKDAY AND WEEKEND;
    x2b2 = (x2b2_0*4 + x2b2_1*3) / 7;

    /* NOTE that even for IDs with ONE dietary recall, there will be predicted
    values for both weekend day and weekday. However, IDs WITH ANY MISSING
    COVARIATES WILL HAVE MISSING x2b2_0 and x2b2_1 */

    * NCI Macro BUG: MISSING VALUES WILL STOP INDIVINT FROM RUNNING SO WE NEED TO
    EXCLUDE MISSING VALUES;
    if x2b2 ne .;
run;
```

Step 4.2: Execute the NCI INDIVINT macro

```
* INCLUDE THE INDIVINT MACRO;
%include "H:\HCHS\NCI_macros\indivint_macro_v1.1.sas" / nosource;

/* NOTE This macro will produce the following error which can be ignored:
Error:  Box-Cox(t,lamt) transformation not selected, so lamt is not applicable. */
%indivint(model12=1, /* Amount model */
          subj1recdata=paramsubj1rec,
          recid=id,
          r24vars=R1 R2,
          min_amt=min_amt,
          var_u1=,
          var_u2=a_var_u2, /* variable created by MIXTRAN macro */
          cov_u1u2=,
          var_e=a_var_e,   /* variable created by MIXTRAN macro */
          lambda=a_lambda, /* variable created by MIXTRAN macro */
          xbeta1=,
          xbeta2=x2b2,
          boxcox_t_lamt=N, /* We don't want predicted value in transformed scale */
          lamt=1,          /* Macro don't run if nothing specified, but 1 is safe*/
          dencalc=y,
          denopt=y,
          u1nlmix=,
          u2nlmix=,
          titles=2,
          notesprt=y);
```

Step 4.3: Saving the output dataset

Since the predicted intake dataset is in the WORK library, we need to save it as a permanent dataset for future use. The variable INDUSINT is the predicted intake (Kipnis and Tooze, personal communication).

```
data pred_energy;
    set _resdata(keep= id indusint);
    rename indusint = pred_energy;
run;

proc sort data=work.subj1rec;    by id; run;
proc sort data=work.pred_energy; by id; run;

data outlib.pred_energy;
 merge work.subj1rec (keep = subjid id r1 r2)
       work.pred_energy;

 by id;

 energy_2dm  = mean(r1,r2);
 diff_energy = pred_energy - energy_2dm;

 rename
  r1 = energy_1
  r2 = energy_2;

  label
  energy_1    = "First recall"
  energy_2    = "Second recall"
  energy_2dm  = "Two-day mean"
  pred_energy = "NCI Predicted energy, kcal";
run;

proc means data=outlib.pred_energy n nmiss min mean max;
 var  energy_1 energy_2 energy_2dm pred_energy diff_energy;
 title "DATA CHECK";
run; title;

title "First 20 records of predicted energy intake";
proc print data = outlib.pred_ENERGY(obs=20); run; title;


/* End of SAS code */
```