**Documentation of the cleaning of the dietary dataset**

*Overview*

The NHANES dietary data from the 11 time periods need to be cleaned and harmonized before being merged into one final dataset. The cleaning of these data is documented in a separate spreadsheet file “Doc\_dietary\_variable\_cleaning\_v1.xlsx”, which lists each variable in each time period and how it is harmonized in the final dataset. This file has 5 sheets, which are described below.

“NHANES\_dietary”: this sheet documents the variable codename, variable description, codename change, and unit change for all variables included in the NHANES dietary recall, that is, variable types that are “survey”, “diet”, “salt”, “water”, “nutrient”, and “fish&shellfish”. The source of these variables is: <https://wwwn.cdc.gov/nchs/nhanes/default.aspx>

“Dietary Fix Category”: this sheet documents, for categorical variables, the category numbers and meanings in each of the 11 time periods, if the category numbers and meanings changed across time. It also documents the harmonized category numbers and meanings, as well as the harmonized variable codenames. This sheet is for the variables in the sheet “NHANES\_dietary”.

“FPED”: this sheet documents the variable codename, variable description, codename change, and unit change for all variables included in the Food Patterns Equivalents Database (FPED), that is, variable type that is “FPED”. It also includes several variables that are of type “survey”. The source of these variables is: <https://www.ars.usda.gov/northeast-area/beltsville-md-bhnrc/beltsville-human-nutrition-research-center/food-surveys-research-group/docs/fped-databases/>

“FPED Fix Category”: this sheet documents, for categorical variables, the category numbers and meanings in each of the 11 time periods, if the category numbers and meanings changed across time. It also documents the harmonized category numbers and meanings, as well as the harmonized variable codenames. This sheet is for the variables in the sheet “FPED”.

“codename\_2day\_combined”: in the original NHANES dietary and FPED data, the 1st day and 2nd day recalls are in separate files, and they have different variable codenames that indicate if it is the 1st day or the 2nd day. For example, the codename for total alcohol intake in the 1st day recall is “DR1TALCO”, while the codename for total alcohol intake in the 2nd day recall is “DR2TALCO”. However, in our final harmonized dataset, we combined the 1st day and 2nd day recalls in one dataset. Thus, we changed both codenames for total alcohol intake to “DRXTALCO”, and added a variable “survey\_day” with a value 1 or 2 to indicate if it is the 1st day or the 2nd day recall. As a result, this sheet documents how the different variable codenames for two days were changed to one common codename.

*Cleaning process*

To clean and harmonize the dietary dataset, we first downloaded the NHANES dietary recall and FPED data, whose 1st day and 2nd day data for different time periods were in separate files. Second, we harmonized the variable codenames, units and categories according to the sheets “NHANES\_dietary”, “Dietary Fix Category”, “FPED”, and “FPED Fix Category”. Note that for NHANES III, the dietary recall data included a first exam (examdr) and a second exam (examdrse), which used different variable codenames. Thus, the variable codenames in the second exam were first harmonized with the first exam as documented in the sheet “NHANES\_dietary”, and then the first exam and second exam data were combined into one dataset for NHANES III. In this NHANES III dataset, a variable “survey\_day” was added with a value 1 or 2 to indicate if it is the first exam or second exam data. Next, this NHANES III dataset was harmonized with the other dietary recall data in NHANES continuous according to the sheets “NHANES\_dietary” and “Dietary Fix Category”.

Third, we merged the 11 time periods together, yielding 4 datasets for NHANES dietary day 1, NHANES dietary day 2, FPED day 1 and FPED day 2. Fourth, we combined the day 1 and day 2 data according to the sheet “codename\_2day\_combined”, yielding 2 datasets for NHANES dietary both days and FPED both days. Finally, we merged these two datasets by the respondent sequence number SEQN\_new to obtain the one final dataset.

*Explanation on the cleaning documentation*

Here we explain in detailed each sheet of the cleaning documentation “Doc\_dietary\_variable\_cleaning\_v1.xlsx”.

Sheets “NHANES\_dietary” and “FPED”: these two sheets have exactly the same structure. Each column is explained below.

* Codename\_original (column A): the original variable codename as downloaded from NHANES and FPED.
* Variable\_description (column B): description of the variable.
* File\_name (column C): the original data file name as downloaded from NHANES and FPED that contains the specified variable.
* File\_summary (column D): brief description of the original data file.
* SDDSRVYR (column E): an integer from -1 to 10 that indicates the NHANES period.
* Year (column F): the year corresponding to the NHANES period.
* Codename\_change (column G): a value 1 indicates that this variable codename was changed to be consistent over time. A blank cell means no change of the variable codename.
* Corrected\_codename\_2day\_separate (column H): the codename that the original codename was changed to. For the same nutrient or food groups intake, the corrected codenames are different between the 1st day recall and the 2nd day recall.
* Codename\_note (column I): there is a note here if the variable is a categorical variable and the categories changed over time. In addition, for NHANES III variables, some specific values such as “8888” represents “not available (NA)”, so here it indicates which values represent NA.
* Unit\_change (column J): a value 1 indicates that this variable’s unit was changed to be consistent over time. A blank cell means no change of the variable unit.
* Old\_unit (column K): if there is a unit change, here indicates the old unit.
* New\_unit (column L): if there is a unit change, here indicates the new unit.
* Unit\_conversion\_factor (column M): if there is a unit change, here indicates the conversion factor between the old unit and the new unit. Values in the old unit need to be **multiplied** by this conversion factor to get values in the new unit.
* Variable\_type (column N): it classifies each variable to one of the 7 types: survey, nutrient, food groups equivalents (FPED), water consumption, salt use, diet, and fish & shellfish consumption.

Sheets “Dietary Fix Category” and “FPED Fix Category”: these two sheets have exactly the same structure. Each column is explained below.

* Codename\_original (column A): the original variable codename as downloaded from NHANES and FPED.
* Categories\_num (column B): the numeric values of the categories for the original codename.
* Categories\_description (column C): the meaning of the numeric values of the categories for the original codename.
* New\_categories (column D): the numeric values of the categories for the harmonized codename.
* New\_categories\_description (column E): the meaning of the numeric values of the categories for the harmonized codename.
* cycle\_-1 to cycle\_10 (columns F to P): an “x” indicates that the specified category number and meaning in columns B and C are available in the corresponding time period (cycle). The cycle name and corresponding period are presented in Table 1.
* corrected\_codename\_2day\_separate (column Q): the codename that the original codename was changed to. For the same nutrient or food groups intake, the corrected codenames are different between the 1st day recall and the 2nd day recall.

Table 1. Cycle name and the corresponding values of the variable “SDDSRVYR” and time periods

|  |  |  |
| --- | --- | --- |
| **Cycle** | **SDDSRVYR** | **NHANES period** |
| Cycle -1 | -1 | 1988-1994 |
| Cycle 1 | 1 | 1999-2000 |
| Cycle 2 | 2 | 2001-2002 |
| Cycle 3 | 3 | 2003-2004 |
| Cycle 4 | 4 | 2005-2006 |
| Cycle 5 | 5 | 2007-2008 |
| Cycle 6 | 6 | 2009-2010 |
| Cycle 7 | 7 | 2011-2012 |
| Cycle 8 | 8 | 2013-2014 |
| Cycle 9 | 9 | 2015-2016 |
| Cycle 10 | 10 | 2017-2018 |

Sheet “codename\_2day\_combined”: as described above, variables from the 1st day and 2nd day recalls were combined and given a new codename, which is documented in this sheet. Each column is explained below.

* Corrected\_codename\_2day\_separate (column A): the 1st day and 2nd day variables codenames after harmonization, corresponding to column H in the sheets “NHANES\_dietary” and “FPED”.
* Variable\_description (column B): description of the variable.
* Survey\_day (column C): indicates if this variable is for 1st or 2nd day recall.
* Final\_codename\_2day\_combined (column D): the new codename that combines the 1st and 2nd day recalls.
* Variable\_type (column E): it classifies each variable to one of the 7 types: survey, nutrient, food groups equivalents (FPED), water consumption, salt use, diet, and fish & shellfish consumption. It corresponds to column N in the sheets “NHANES\_dietary” and “FPED”.