

**Big Data and machine learning-based nutritional and health assessment of fatty acids**

PI Olivier Jolliet and Dr. Lei Huang,   
with support from Dr. Vy Nguyen, Neil Zhao and De Eliseu Verly

University of Michigan, School of public Health, Environmental Health Sciences

(v2, 23 March 2022)

**Documentation of the dietary dataset**

*Overview*

The dietary data were collected from the National Health and Nutrition Examination Survey (NHANES) designed by the United States Centers for Disease Control and Prevention (CDC), which is a cross-sectional, ongoing population-based study to ascertain the nutritional and health status of children and adults in the US.

The dietary dataset covered the period from 1988 to 2018, which includes 11 sub-periods. The 1988-1994 period corresponds to NHANES III, while the 1999-2018 period corresponds to NHANES continuous, with 10 two-year cycles. For example, year 1999-2000 corresponds to the 1st cycle of NHANES continuous, and year 2001-2002 corresponds to the 2nd cycle of NHANES continuous, etc.

The dietary dataset includes a total of 117181 unique respondents with non-missing data. The dietary intake data were obtained using 24-hr dietary recall questionnaire. 117066 respondents have the 1st day dietary recall, while 62486 have the 2nd day recall. Among these, 62372 respondents have both the 1st day and 2nd day recalls. Thus, 54694 respondents only have the 1st day recall data, while 114 respondents only have the 2nd day recall data. The number of respondents in each time period is listed in Table 1. We have also estimated the usual intakes based on the 2-day or 1-day dietary recall data, which are included in this dataset.

Table 1. Number of respondents in each NHANES period



*Dataset structure*

For the structure of the dietary dataset, each row represents one 24-hr recall of one respondent, and each column represents one variable. Each respondent is identified by a unique sequence number with the variable name “SEQN”. Thus, for respondents with 1st day or 2nd day recall, there are 2 rows for each respondent, one row for the recall and the other row for the estimated usual intake. For respondents with both 1st day and 2nd day recalls, there are 3 rows for each respondent – two rows for the 1st and 2nd day recalls and one row for the estimated usual intake. There are a total of 324,374 rows in the dataset, including respondents with non-missing and missing data. A screenshot of the dataset is presented in Figure 1.

Table

Description automatically generated

Figure 1. Screenshot of the final dietary dataset.

*Variable documentation*

There are a total of 317 variables in the dataset. Detailed information of each variable is provided in a separate spreadsheet document “Doc\_dietary\_variable\_list.xlsx”. In this spreadsheet:

“Variable name” (column A) presents the codename of each variable.

“Variable description” (column B) describes the meaning of each variable.

“Variable type” (column C) classifies each variable to one of 7 types: survey, nutrient, food groups equivalents (FPED), water consumption, salt use, diet, and fish & shellfish consumption. The number of each variable type is listed in Table 2.

“Availability” (columns D to N) indicates if each variable is available in each of the 11 NHANES periods. An “x” indicates that the variable is available in the specified period, while a blank cell indicates that the variable is not available. Certain variables are noted as “unit change”, for example, the variable “DRXTVARA” in NHANES III, which means that this variable was recorded in the unit of retinol equivalent (RE) in NHANES III that was changed to the unit of retinol activity equivalent (RAE) to be consistent with the later years. Certain variables are noted as “calculated”, for example, the variable “DRXT\_V\_REDOR\_TOTAL” in NHANES III, which means that this variable did not exist in the original NHANES III data, but was calculated from other variables. The variable “DRXT\_V\_REDOR\_TOTAL” was calculated as the sum of “DRXT\_V\_REDOR\_OTHER” and “DRXT\_V\_REDOR\_TOMATO”.

“Data source and documentation” (columns O to R) provides the source of the raw data and the link of the NHANES documentations. Columns O an P provide the file description and documentation link for NHANES III (1988-1994), while columns Q and R provide the file description and documentation link for NHANES continuous (1999-2018).

Table 2. Number of variables by type

|  |  |
| --- | --- |
| **Variable type** | **Count** |
| survey | 31 |
| water | 5 |
| salt | 6 |
| diet | 15 |
| fish & shellfish | 66 |
| nutrient | 157 |
| FPED | 37 |
| Mortality | 7 |

*Variable explanation*

Explanations on several important variables are as follows:

SEQN: the original respondent sequence number reported in NHANES.

SEQN\_new: since there are duplicate sequence numbers in NHANES III and NHANES continuous, we created this new sequence number that added a prefix to the original sequence number. A prefix “C-“ was added to the sequence numbers in NHANES continuous, while a prefix “I-“ was added to those in NHANES III.

SDDSRVYR: this is a number that indicates the NHANES period, as listed in Table 3.

RIDAGEYR: age of the respondent in years. It equals zero if the respondent is an infant less than 12 months old.

RIAGENDR: gender of the respondent. 1 is male, 2 is female.

survey\_day: a value of 1 indicates this is the 1st day recall of the respondent; a value of 2 indicates this is the 2nd recall; a value of 0 indicates this is the estimated usual intake.

Description of the other variables can be found in the file “Doc\_dietary\_variable\_list.xlsx”.

Table 3. Values of the variable “SDDSRVYR” and the corresponding time periods

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

Table 4 presents the dietary variables that are available in more than 5 NHANES periods. The full list of variable is decsribed in the excel document "Doc\_dietary\_variable\_list\_v3a".

Table 4. Dietary variables available in more than 5 NHANES periods

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable name** | **Variable description** | **Variable type** | **Number of cycles covered** | **Period covered** |
| SEQN\_new | Respondent sequence number that includes an identifier for NHANES III and NHANES continuous | survey | 11 | 1988-2018 |
| SEQN | Respondent sequence number | survey | 11 | 1988-2018 |
| SDDSRVYR | Sample cycle | survey | 11 | 1988-2018 |
| RIDAGEYR | Age in years | survey | 11 | 1988-2018 |
| RIAGENDR | Gender | survey | 11 | 1988-2018 |
| survey\_day | Dietary survey day (day 1 or day 2 or usual intake) | survey | 11 | 1988-2018 |
| WTDRD1 | Dietary day one 2-Year sample weight | survey | 10 | 1999-2018 |
| DRXEXMER | Interviewer ID code (sequential #) | survey | 10 | 1999-2018 |
| DRXDAY | Intake day of the week | survey | 11 | 1988-2018 |
| DRXLANG | Language SP/Proxy used mostly | survey | 11 | 1988-2018 |
| DRXDRSTZ | Dietary recall status | survey | 11 | 1988-2018 |
| DRXMRESP | Who was the main respondent for this interview? | survey | 11 | 1988-2018 |
| DRXTNUMF | Number of foods | survey | 9 | 2001-2018 |
| WTDRXD | Dietary two-day sample weight | survey | 8 | 2003-2018 |
| DRABF | Breast-fed infant (either day) | survey | 8 | 2003-2018 |
| DRDINT | Number of days of intake | survey | 8 | 2003-2018 |
| DRXHELP | Who helped in responding for this interview | survey | 8 | 2003-2018 |
| DRXDBIH | # of days b/w intake and HH interview | survey | 6 | 2007-2018 |
| DRXTKCAL | Energy (kcal) | nutrient | 11 | 1988-2018 |
| DRXTPROT | Protein (gm) | nutrient | 11 | 1988-2018 |
| DRXTCARB | Carbohydrate (gm) | nutrient | 11 | 1988-2018 |
| DRXTTFAT | Total fat (gm) | nutrient | 11 | 1988-2018 |
| DRXTSFAT | Total saturated fatty acids (gm) | nutrient | 11 | 1988-2018 |
| DRXTMFAT | Total monounsaturated fatty acids (gm) | nutrient | 11 | 1988-2018 |
| DRXTPFAT | Total polyunsaturated fatty acids (gm) | nutrient | 11 | 1988-2018 |
| DRXTCHOL | Cholesterol (mg) | nutrient | 11 | 1988-2018 |
| DRXTFIBE | Dietary fiber (gm) | nutrient | 11 | 1988-2018 |
| DRXTVARA | Vitamin A (RAE) | nutrient | 11 | 1988-2018 |
| DRXTVB1 | Thiamin (Vitamin B1) (mg) | nutrient | 11 | 1988-2018 |
| DRXTVB2 | Riboflavin (Vitamin B2) (mg) | nutrient | 11 | 1988-2018 |
| DRXTNIAC | Niacin (mg) | nutrient | 11 | 1988-2018 |
| DRXTVB6 | Vitamin B6 (mg) | nutrient | 11 | 1988-2018 |
| DRXTFOLA | Total Folate (mcg) | nutrient | 11 | 1988-2018 |
| DRXTVB12 | Vitamin B12 (mcg) | nutrient | 11 | 1988-2018 |
| DRXTVC | Vitamin C (mg) | nutrient | 11 | 1988-2018 |
| DRXTATOC | Vitamin E (ATE) (mg) | nutrient | 11 | 1988-2018 |
| DRXTCALC | Calcium (mg) | nutrient | 11 | 1988-2018 |
| DRXTPHOS | Phosphorus (mg) | nutrient | 11 | 1988-2018 |
| DRXTMAGN | Magnesium (mg) | nutrient | 11 | 1988-2018 |
| DRXTIRON | Iron (mg) | nutrient | 11 | 1988-2018 |
| DRXTZINC | Zinc (mg) | nutrient | 11 | 1988-2018 |
| DRXTCOPP | Copper (mg) | nutrient | 11 | 1988-2018 |
| DRXTSODI | Sodium (mg) | nutrient | 11 | 1988-2018 |
| DRXTPOTA | Potassium (mg) | nutrient | 11 | 1988-2018 |
| DRXTSELE | Selenium (mcg) | nutrient | 11 | 1988-2018 |
| DRXTCAFF | Caffeine (mg) | nutrient | 11 | 1988-2018 |
| DRXTTHEO | Theobromine (mg) | nutrient | 10 | 1999-2018 |
| DRXTALCO | Alcohol (gm) | nutrient | 11 | 1988-2018 |
| DRXTMOIS | Moisture (gm) | nutrient | 11 | 1988-2018 |
| DRXTS040 | SFA 4:0 (Butanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS060 | SFA 6:0 (Hexanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS080 | SFA 8:0 (Octanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS100 | SFA 10:0 (Decanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS120 | SFA 12:0 (Dodecanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS140 | SFA 14:0 (Tetradecanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS160 | SFA 16:0 (Hexadecanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTS180 | SFA 18:0 (Octadecanoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTM161 | MFA 16:1 (Hexadecenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTM181 | MFA 18:1 (Octadecenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTM201 | MFA 20:1 (Eicosenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTM221 | MFA 22:1 (Docosenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP182 | PFA 18:2 (Octadecadienoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP183 | PFA 18:3 (Octadecatrienoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP184 | PFA 18:4 (Octadecatetraenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP204 | PFA 20:4 (Eicosatetraenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP205 | PFA 20:5 (Eicosapentaenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP225 | PFA 22:5 (Docosapentaenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTP226 | PFA 22:6 (Docosahexaenoic) (gm) | nutrient | 11 | 1988-2018 |
| DRXTSUGR | Total sugars (gm) | nutrient | 9 | 2001-2018 |
| DRXTRET | Retinol (mcg) | nutrient | 10 | 1988-1994, 2001-2018 |
| DRXTACAR | Alpha-carotene (mcg) | nutrient | 9 | 2001-2018 |
| DRXTBCAR | Beta-carotene (mcg) | nutrient | 10 | 1988-1994, 2001-2018 |
| DRXTCRYP | Beta-cryptoxanthin (mcg) | nutrient | 9 | 2001-2018 |
| DRXTLYCO | Lycopene (mcg) | nutrient | 9 | 2001-2018 |
| DRXTLZ | Lutein + zeaxanthin (mcg) | nutrient | 9 | 2001-2018 |
| DRXTFA | Folic acid (mcg) | nutrient | 9 | 2001-2018 |
| DRXTFF | Food folate (mcg) | nutrient | 9 | 2001-2018 |
| DRXTFDFE | Folate, DFE (mcg) | nutrient | 9 | 2001-2018 |
| DRXTVK | Vitamin K (mcg) | nutrient | 9 | 2001-2018 |
| DRXTATOA | Added alpha-tocopherol (Vitamin E) (mg) | nutrient | 8 | 2003-2018 |
| DRXTB12A | Added vitamin B12 (mcg) | nutrient | 8 | 2003-2018 |
| DRXTCHL | Total choline (mg) | nutrient | 7 | 2005-2018 |
| DRXTVD | Vitamin D (D2 + D3) (mcg) | nutrient | 7 | 1988-1994, 2007-2018 |
| DRXT\_G\_TOTAL | Total number of grain ounce equivalents | FPED | 11 | 1988-2018 |
| DRXT\_G\_WHOLE | Number of whole grain ounce equivalents | FPED | 11 | 1988-2018 |
| DRXT\_G\_REFINED | Number of non-whole grain ounce equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_TOTAL | Total number of vegetable cup equivalents, excl legumes | FPED | 11 | 1988-2018 |
| DRXT\_V\_DRKGR | Number of dark-green vegetable cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_REDOR\_OTHER | Number of orange vegetable cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_STARCHY\_POTATO | Number of white potato cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_STARCHY\_OTHER | Number of other starchy vegetable cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_REDOR\_TOMATO | Number of tomato cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_V\_OTHER | Number of other vegetable cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_F\_TOTAL | Total number of fruit cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_F\_CITMLB | Number of citrus, melon, berry cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_F\_OTHER | Number of other fruit cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_D\_TOTAL | Total number of milk group (milk, yogurt & cheese) cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_D\_MILK | Number of milk cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_D\_YOGURT | Number of yogurt cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_D\_CHEESE | Number of cheese cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_PF\_MPS\_TOTAL | Oz cooked lean meat from meat, poultry, fish | FPED | 11 | 1988-2018 |
| DRXT\_PF\_MEAT | Oz cooked lean meat from beef, pork, veal, lamb, and game | FPED | 11 | 1988-2018 |
| DRXT\_PF\_ORGAN | Oz cooked lean meat from organ meats | FPED | 11 | 1988-2018 |
| DRXT\_PF\_CUREDMEAT | Oz cooked lean meat from franks, sausages, luncheon meats | FPED | 11 | 1988-2018 |
| DRXT\_PF\_POULT | Oz cooked lean meat from chicken, poultry, and other poultry | FPED | 11 | 1988-2018 |
| DRXT\_PF\_SEAFD\_HI | Oz cooked lean meat from fish, other seafood high in Omega-3 | FPED | 11 | 1988-2018 |
| DRXT\_PF\_SEAFD\_LOW | Oz cooked lean meat from fish, other seafood low in Omega-3 | FPED | 11 | 1988-2018 |
| DRXT\_PF\_EGGS | Oz equivalents of lean meat from eggs | FPED | 11 | 1988-2018 |
| DRXT\_PF\_SOY | Oz equivalents of lean meat from soy product | FPED | 11 | 1988-2018 |
| DRXT\_PF\_NUTSDS | Oz equivalents of lean meat from nuts and seeds | FPED | 11 | 1988-2018 |
| DRXT\_V\_LEGUMES | Number of cooked dry beans and peas cup equivalents | FPED | 11 | 1988-2018 |
| DRXT\_OILS | Grams of discretionary Oil | FPED | 11 | 1988-2018 |
| DRXT\_SOLID\_FATS | Grams of discretionary Solid fat | FPED | 11 | 1988-2018 |
| DRXT\_ADD\_SUGARS | Teaspoon equivalents of added sugars | FPED | 11 | 1988-2018 |
| DRXT\_A\_DRINKS | Total drinks of alcohol | FPED | 11 | 1988-2018 |
| DRXT\_F\_JUICE | Fruit juices, citrus and non citrus (cup eq.) | FPED | 7 | 2005-2018 |
| DRXT\_V\_REDOR\_TOTAL | Total red and orange vegetables (tomatoes + other red and orange) (cup eq.) | FPED | 11 | 1988-2018 |
| DRXT\_V\_STARCHY\_TOTAL | Total starchy vegetables (white potatoes + other starchy) (cup eq.) | FPED | 11 | 1988-2018 |
| DRXT\_PF\_LEGUMES | Legumes computed as protein foods (oz. eq.) | FPED | 7 | 2005-2018 |
| DRXT\_PF\_TOTAL | Total meat, poultry, seafood, organ meats, cured meat, eggs, soy, and nuts and seeds; excludes legumes (oz. eq.) | FPED | 11 | 1988-2018 |
| DRX.300 | Compare food consumed yesterday to usual | diet | 11 | 1988-2018 |
| DRQSDIET | On special diet? | diet | 8 | 2003-2018 |
| DRQSDT1 | Weight loss/low cal/low carb/hi pro diet | diet | 8 | 2003-2018 |
| DRQSDT2 | Low fat/ Low cholesterol diet | diet | 8 | 2003-2018 |
| DRQSDT3 | Low salt/ Low sodium diet | diet | 8 | 2003-2018 |
| DRQSDT4 | Sugar free/ low sugar diet | diet | 8 | 2003-2018 |
| DRQSDT5 | Low fiber diet | diet | 7 | 2005-2018 |
| DRQSDT6 | High fiber diet | diet | 8 | 2003-2018 |
| DRQSDT7 | Diabetic diet | diet | 8 | 2003-2018 |
| DRQSDT8 | Weight gain/ Muscle building diet | diet | 8 | 2003-2018 |
| DRQSDT91 | Other special diet | diet | 8 | 2003-2018 |
| DRQSDT9 | Low carbohydrate diet | diet | 6 | 2007-2018 |
| DRQSDT10 | High protein diet | diet | 6 | 2007-2018 |
| DRX.320Z | Total plain water drank yesterday (gm) | water | 11 | 1988-2018 |
| DRX.330Z | Total home tap water drank yesterday(gm) | water | 10 | 1999-2018 |
| DRXBWATZ | Total bottled water drank yesterday (gm) | water | 8 | 2003-2018 |
| DRXTWSZ | Tap Water Source | water | 8 | 2003-2018 |
| DBD100 | How often add salt to food at table | salt | 10 | 1988-1994, 2001-2018 |
| DBQ095Z | What type of salt {do you/does SP} usually add to {your/his/her/SP's} food at the table? | salt | 10 | 1988-1994, 2001-2018 |
| DRQSPREP | Salt used in preparation? | salt | 8 | 2003-2018 |

**Combination with other dataset**

The dietary data were merged with different background dataset that were already collected and cleaned at University of Michigan in the frame of other projects:

Mortality data are from the National Death Index (NDI), a database of all deaths in the US (variables in Table 5). The mortality information includes mortality status at follow-up, duration of follow-up, and cause of death. These mortality data are liked to NHANES via the respondent sequence number (SEQN). NHANES participants are followed up until Dec. 31, 2015.

Table 5. Mortality variable available from 1988-2014

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ELIGSTAT | Eligibility Status for Mortality Follow-up | mortality | 9 | 1988-2014 |
| MORTSTAT | Final Mortality Status | mortality | 9 | 1988-2014 |
| UCOD\_LEADING | Underlying Leading Cause of Death: Recode | mortality | 9 | 1988-2014 |
| DIABETES | Diabetes Flag from Multiple Cause of Death | mortality | 9 | 1988-2014 |
| HYPERTEN | Hypertension Flag from Multiple Cause of Death | mortality | 9 | 1988-2014 |
| PERMTH\_INT | Number of Person Months of Follow-up from NHANES interview date | mortality | 9 | 1988-2014 |

Demographics and questionnaire data are taken from the NHANES dataset curated at University of Michigan. The following physiological measurements and response indicators are also made available (Table 6).

Table 6. Physiological indicators available to the project

