**Intake24 dietary data cleaning**

This document provides an overview of principles of checking and cleaning Intake24 dietary intake data. It is intended as a guide to assist researchers with developing their own QA/QC protocol, which should be informed by the individual study objectives, outcomes and data analysis plan. It is also recommended that this document is read in conjunction with the following guidance:

* [Coding Intake24 missing foods](https://www.mrc-epid.cam.ac.uk/wp-content/uploads/2025/03/Missing-foods-protocol-for-external-users_v3.0.docx)
* [Intake24 data dictionary](https://www.mrc-epid.cam.ac.uk/research/measurement-platform/dietary-assessment/intake24/) 2023 - specific to the 'STANDARD\_2023' scheme which is the default scheme (set of questions asked throughout the recall)

There are a number of checks that can be carried out on Intake24 dietary data. The type and frequency of checks, as well as any decisions on adjusting or excluding data will vary for different studies. Not all checks will be required for all studies. Suggested checks are listed below, along with the overall purpose and description.

1. **Basic monitoring checks**
2. **Box plots high energy recalls**
3. **Checks for low energy recalls**
4. **Box plots portion size**
5. **Box plots nutrients**
6. **Detailed quality checks**

To manage data cleaning processes, creating a database (e.g. Access) may be desirable, especially for larger datasets.

1. **Basic monitoring checks**

**Basic monitoring checks may be useful to be performed at the start of a study, and/or at agreed intervals/frequency.**

**The overall aim of these checks is to monitor incidence of specific dietary data issues as an indicator of quality and to provide a baseline and an alert to investigate further if above an expected threshold.**

The standard basic monitoring checks may include:

1. Count of missing foods
2. Count of recalls with less than 10 items (excluding associated foods1)
3. Count of recalls that took 2 minutes or less to complete
4. Count of recalls that contain less than 400kcals
5. Count of recalls that contain over 4000kcals

Standard checks and the cut-offs of individual checks (e.g. recall completion time or high and low energy intake) may be modified according to the need of specific studies/populations and in response to any issues emerging in the dietary data.

1. **Box plots high energy recalls**

**The overall aim of this check is to identify implausibly high energy intakes for a recall that may require adjusting or recommending excluding in the dietary dataset. This check may also identify any systematic issues.**

Box plots can be generated using any statistical package. An example of code using R and the output is provided in Appendix 1. Plots are commonly generated using the 3\*IQR (interquartile range) rule to detect outliers, however, this range can be changed as required for your study.

Using the box plots identify and investigate outliers:

Step 1: Complete an initial visual inspection of the box plots to identify and flag highly implausible total energy intakes:

* Use your own knowledge of reasonable daily total energy intakes, according to age and sex etc.
* Use any previous findings or other research available to you and relevant to your population / study
* Take into account the distance an outlier is from the end of the distribution tail – the further away they are the more influential they will be on data summaries

Step 2: A second review by another researcher to confirm, reject or expand the list of outliers for further investigation

Step 3: Investigate the agreed outliers

Total energy intake outliers can be reviewed in the Intake24 data extract: identify and document any specific food or drinks items in the recall that may be contributing to the extreme energy. Decide on whether to adjust foods or drinks that are contributing to high energy intake or to exclude the recall from your dataset – during this step, you may wish to refer to guidelines outlined by the National Cancer Institute: <https://epi.grants.cancer.gov/asa24/resources/ASA24_2024_Data_Review_and_Cleaning.pdf>.

Total energy intake box plots may also provide an indication of any systematic issues.

Prior to making any final decisions to adjust or exclude recalls, portion size box plots checks may help inform decisions: see #4 below.

1. **Checks for low energy Recalls**

**The overall aim of this check is to consider recalls with very low energy intakes, which may indicate incomplete dietary reporting and may require flagging, adjusting or excluding in the dietary dataset. This check may also identify any systematic issues.**

As a guide recalls less than 400kcals can be identified in the dataset for further investigation.

Check whether there is a valid reason for a low intake e.g. the participant reported feeling unwell.

Decisions on whether to exclude or adjust low energy recalls will depend on individual study requirements.

1. **Box plots portion size**

**The overall aims of box plots portion checks are**

* **to identify and document highly implausible large portion sizes**
* **consider adjustment or recommend exclusion of recalls in the dietary data**
* **identify any systematic issues**

Generate and examine box plots as in #2 above, but use the portion size variable in the Intake24 data extract. Create box plots for each food group, for example NDNS sub food groups, or key food groups, or Intake24 or other categorisation, see [Intake24 data dictionary](https://www.mrc-epid.cam.ac.uk/research/measurement-platform/dietary-assessment/intake24/).

Using the box plots identify and investigate outliers:

Step 1: Complete an initial visual review of the box plots to identify and flag highly implausible large portion sizes, apply the following:

* Use your own knowledge of reasonable portions consumed, according to age and sex etc.
* Using previous findings from data checks and other relevant datasets
* Taking into account the distance an outlier is from the end of the distribution tail – the further away they are the more influential they will be on data summaries

Step 2: A second review by another researcher to confirm, reject or expand the list of outliers

Step 3: Investigate the agreed outliers

The details of the portion outliers can be reviewed in the Intake24 data extract. Examine portion outliers in the context of all the food and drink items in the recall to determine if an adjustment or excluding the recall from the dataset is necessary. Portion size box plots may also provide an indication of any systematic issues.

**Adjustments or excluding recalls as a result of energy or portion size outliers**

Decisions to adjust or exclude a recall are not always clear cut, and are likely to be influenced by the specific study, research objectives, time and resources. The frequency of the occurrence, size and scale of the outlier and its impact on the data overall will also inform adjustments or excluding recalls.

1. **Box plots nutrients**

**The overall aim of this check is to examine nutrient variables for extreme outliers. This may indicate potential error(s) in the Nutrient Databank food code(s) (e.g. a food composition error) or a data entry error in a recall not yet identified through energy and portion size box plots checks. It is also a reasonableness check to identify any potential issues with food entries at a day recall level.**

This check is an option that can be applied at the end of a study, when all dietary data has been collected.

Generate and examine box plots as in #2 and #4 above, but use the recall level nutrient variables in the Intake24 data extract. Generally the nutrient box plots should NOT include supplements from the Intake24 data extract (exclude NDNS food group 54). Inclusion of supplements may mask extreme nutrient outliers.

Where outliers are found, identify which recalls(s) contain the particular outlier nutrient and then check which food(s) are providing the greatest amount of this outlier nutrient. This may highlight an extreme portion size as the cause of the outlier or it may be a food composition error in a food code. If it is a food code error, you will need to adjust this in your dataset. It will also need to be corrected in Intake24 databases, please notify us at: [Intake24.support@mrc-epid.cam.ac.uk](mailto:Intake24.support@mrc-epid.cam.ac.uk).

1. **Detailed quality checks**

**The overall aims of the detailed checks are to monitor performance of Intake24 in a particular study population and/or in a new setting, and to pick up issues that may need to be addressed in the tool or in the placement of the tool in the study for example, you may wish to review or update your fieldwork instructions or guidance to study participants. These checks may also be useful to support fieldwork monitoring particularly at the start of a study or if there has been a change of fieldwork protocol.**

Detailed quality checks can be very time consuming and are not recommended as a frequent check on the whole dataset. For example, these checks may be helpful to be carried out at agreed intervals on a randomly selected 10% sample of the dietary data to understand potential issues in the dietary data. For larger or smaller studies the percentage may be set differently, depending on timelines or resources. There are three components to the detailed quality checks:

1. Multiple foods on one line
2. Orphan foods
3. Inconsistencies
4. **Multiple foods on one line**

‘Multiple foods on one line’ (MF1L) and ‘Multiple foods on one line not fully coded’ includes all occasions where a participant may have entered multiple foods on one line as the search term (e.g. `*chips peas beans`*) and potentially omitted. Note that such multiple foods may subsequently be entered later in the recall through associated food prompts or the participant may enter the foods separately later in the recall.

1. **Orphan foods**

‘Orphan foods’ are reported foods where the name of the dish entered in the search term does not match the food name ultimately recorded by the participant (e.g. `*Beef Ramen*` typed in the search term and only `*beef steak*` recorded).

Orphan foods may also be where it could reasonably be assumed that other foods would generally accompany them (e.g. if someone only recorded steak for dinner and no other accompaniments), however this cannot be known for certain. Dishes such as lasagne and shepherd’s pie are not counted as orphan foods as participants may be more likely to eat these types of dishes on their own.

1. **Inconsistencies**

‘Inconsistencies’ are where the participant appears (from the search term) to have searched for one food or drink item, and ultimately recorded something different to the search term (e.g. searched for beef and recorded prawn). However, only the initial search term is retained in the data output and a participant can over-type/redo their search for a completely different food, so the inconsistency may be legitimate.

Please do get in touch with us at: [Intake24.support@mrc-epid.cam.ac.uk](mailto:Intake24.support@mrc-epid.cam.ac.uk) if you have any queries.

**Appendix 1: Box plots production code**

The R code below can be used to create boxplots of daily total energy intake by age and sex group (18-39yrs, 40-59yrs and 60+yrs – Female and Male).

If you wish to create nutrient box plots, please replace ‘Energykcal’ with the relevant nutrient variable.

If you wish to create portion size box plots, produce a dataset which contains the portion size of each food consumed within the food group of interest first and then amend the R code below as necessary.

Dataset name e.g.: Intake24data

Variables: Energykcal (numeric), Age (numeric), Sex (categorical F/M)

***R Syntax***

# create factor for Age and Sex

*Intake24data$AgeSex.group[Intake24data$Age>=18 & Intake24data$Age<=39 & Intake24data$Sex=="F"] = "18-39yrs Female"*

*Intake24data$AgeSex.group[Intake24data$Age>=18 & Intake24data$Age<=39 & Intake24data$Sex=="M"] = "18-39yrs Male"*

*Intake24data$AgeSex.group[Intake24data$Age>=40 & Intake24data$Age<=59 & Intake24data$Sex=="F"] = "40-59yrs Female"*

*Intake24data$AgeSex.group[Intake24data$Age>=40 & Intake24data$Age<=59 & Intake24data$Sex=="M"] = "40-59yrs Male"*

*Intake24data$AgeSex.group[Intake24data$Age>=60 & Intake24data$Sex=="F"] = "60+yrs Female"*

*Intake24data$AgeSex.group[Intake24data$Age>=60 & Intake24data$Sex=="M"] = "60+yrs Male"*

*Intake24data$AgeSex.group = as.factor(Intake24data$AgeSex.group)*

# load the ‘car’ library which is required to produce the boxplots

*library(car)*

*Boxplot(Energykcal ~AgeSex.group), data=Intake24data,*

*main = "Energy (kcal/day)", ylab="Energy (kcal/day)", xlab="Age/Sex group",*

*range=3, id.n=10, pch=16)*

# This will produce a boxplot with the row number which contains the identified outlier labelled # on the boxplot and these row numbers will be printed in the R console

# The ‘*range=*’ option determines how far from the nearest quartile an outlier is identified

# *range=3* specifies that any value more than 3xIQR (interquartile range) from the nearest

# quartile (25% or 75% centiles) is marked as an outlier

# the ‘*id.n=*’ option determines how many outliers are presented for each factor level on

# the x-axis

